SESSION 1C

FUTURE INTERNET
Towards Resilient Networks Using Situation Awareness

M Luxia Liu
College of Computer and Communication
Lanzhou University of Technology
Lanzhou, China
m.liu@lancaster.ac.uk

David Hutchison
School of Computing and Communication
Lancaster University
Lancaster, UK
d.hutchison@lancaster.ac.uk

Abstract—Resilience will be an essential design and operational characteristic of future networks. This paper first surveys resilient networks and Situation Awareness (SA) in cyberspace. Then, SA is introduced in the detection module development of our dynamic adaptation architecture that realizes the resilience control loop. Finally, a state transition model based on Petri net analyzing resilience is presented. Further detailed development including a fusion algorithm to analyze failure based on SA and state transition according to failure model evaluating resilience forms the basis for further work.

Keywords—resilience; situation awareness; resilience analysis; failure model

I. INTRODUCTION

Resilience, which is defined as the ability of the networks to provide and maintain an acceptable level of service in the face of various faults and challenges to normal operation, must be viewed as an essential design and operational characteristic of future networks in general, and the Global Internet in particular [1]. To achieve information resilience, it is necessary for cyber defense to move from a “reactive” paradigm to a “proactive” one [2]. In resilient networks, to anticipate failure and measure resilience, it is necessary to know the results as follows. Where did the failure happen? Are the services impacted? Can the system recover to a normal state or provide degraded service? To answer these questions, SA (Situation Awareness) is studied in this paper because SA requires an understanding of current activities, an ability to anticipate what may happen next, and techniques to analyze the threat or impact of current activities and predictions [3].

Resilient networks as an infrastructure provided to customers need to confirm the service level applicants require, need threat and challenge models, need to quantify the service requirements and operational state, need to manage the operational state, need to recover to acceptable service after failure occurs. It is obvious that the rapid growth of the Internet and the World Wide Web makes the current networks more vulnerable, diverse and heterogeneous. Therefore, improving the resilience ability through situation awareness in the heterogeneous networks environment is an emergent problem. In light of the foregoing, this paper discusses the related work of resilience and SA in cyber domain and provides a framework of dynamic detection module of our resilience control loop for understanding and analyzing resilience based on state transition.

II. RELATED WORK REGARDING RESILIENCE

Resilience is a semantically overloaded term in the sense that it means somewhat different things in different fields, such as ecological resilience, economics and business resilience, industrial and organizational resilience, networks resilience, psychological resilience and socio-ecological resilience [4]. In resilience engineering, Madni [4] argued that resilient system must have the ability to anticipate, learn and adapt. In the process of anticipation, environment can be changed expected which will promote desirable outcomes and circumvent disruptions in the future. Learning encompasses continuously monitoring the external environment and system operation for contextually-relevant changes and reflecting them in the system’s models and knowledge bases. Adaptation involves the ability to adjust to new circumstances through reconfiguration and dynamic re-optimization of available capacity and resources. It is necessary for resilient systems to collaborate and work as a team while ensuring awareness of the work processes at all levels with efficient communication [5].

In the early time, networks resilience was as probabilistic measure of potential disconnections in a network [6, 7]. Resilience mechanism and resilience analysis work in case of a network connecting failure. For example, Joseph [8] provides a method for resilient and reliable end-to-end connectivity in heterogeneous networks. Cholda [9] made the first extensive survey of ideas related to differentiation of communication services with respect to resilience, which needs to provide services with different resilience characteristics such as the availability, continuity and duration of downtimes in the same network. Menth [10] presents a framework for resilience analysis of packet-switched communication networks about ingress-egress unavailability and link congestion due to failures, changes of user behavior and changed interdomain routing. Rerouting in networks is considered as resilience for networks failure [11, 12]. These works considering networks failure of resilience are usually applied by lower layers.

Our resilience in communication networks is designed to meet the multilevel bottom-up requirements from physical channel, links and nodes, topology, path, Internet working, transport to application. The ResiliNets strategy is shown in Fig. 1 [1]. And then, a systematic approach to building resilient
networked systems in our work is shown in Fig.2 [13]. It needs automatic connecting between different components rather than simple combination of technology.

III. SITUATION AWARENESS IN CYBER DOMAIN

Bass first presented multisensor data fusion for distributed IDS, and gave the models of cyberspace SA [15] [16]. Salema and Tadda [17, 3, 18, 19] drew further on the JDL model [20] and Endsley’s models [14] to bridge the gap and develop a common framework for SA as shown in Fig.3. They rely on the JDL model for levels 0, 1 and 4 but utilize Endsley’s notions of Perception, Comprehension and Projection (which they refer to as Anticipation) as the overarching framework. In addition, they have added an initial data requirements component to enable top down control over the entire process. Gong [21] presented cyberspace SA research framework, but has no details of their model. The framework is based on JDL and Endsley’s model giving the content of Cyber SA research. Current researches focus on Cyber SA model [22], knowledge representation [23] and assessment method [24].

SA for cyber defense can be viewed as a three–phase process and consists of at least seven aspects [25]: situation recognition (including being aware of the current situation and the quality of the collected situation awareness information items and the knowledge-intelligence-decisions derived from these information items); situation comprehension (including being aware of the impact of the attack, actor (adversary) behavior, and why and how the current situation is caused); situation projection (including being aware of how situations evolve, and assessing plausible futures of the current situation). Existing approaches to gain cyber situation-awareness consist of vulnerability analysis (using attack graphs), intrusion detection and alert correlation, attack trend analysis, causality decision superiority, can meet the dynamical resilience requirements.
According to the level model, from the perspective of data fusion, logged events data is considered level 0 and early level 1. Fusing the sensed data into level 1 information typically consists of either categorizing the alert or event based on the attack step it represents, or correlating multiple alerts together into an attack step, or both. This activity is essentially what the alert correlation systems [29]. Identifying the list of potential tracks is determined by matching the attack steps within a track to a priori patterns or models which is indicative of level 2 data fusion. This matching can be performed by various algorithms such as Bayesian Networks, Markov models, or graph theory [30]. It’s important to note that so far there have not been included any information regarding the network topology, network services, or the missions or importance of the computers within the networks [3]. To estimate resilience of networks, many factors will be considered in the process of perception including topology, traffic, resource, security, fault and survival.

IV. HOW TO USE SA IN RESILIENT NETWORKS

From all the frameworks or models of SA, it is clear that SA can be used for realizing resilience dynamically. Therefore, we developed a model of Resilience through Situation Awareness (RSA) based on our dynamic adaptation architecture. And then we analyze the factors that affect the resilience and will define potential situation that we are looking for in the first level. Such situation can be divided into attacks and failures or interruptions. Once we have defined a set of situations that we are interested in and the set of data sources available we will store the necessary data and the processes required by each source to translate and cleanse the data. According to the RSA model, we can analyze and anticipate the challenges affecting resilience. After identifying the challenges by fusing the information we obtained, we need to define the state transition and recovery. The most important here is to estimate the resilient state based on the state transition model.

A. Model of RSA

To develop the detection module of our dynamic adaptation architecture realizing the resilience control loop, systematical resilience detection model based on SA with regard to perception, comprehension, and anticipation is shown in Fig.4. The result of detection will be stored in DISco (Distributed Store for challenges and their outcome) [13], according to which, networks resilience manager can make decision using our remediation and recovery. When we take measures by remediation, the state of networks described by \( S_{00} \), \( S_{ij} \) to \( S_{nn} \) will change from one to another.

In the process of perception of our model, it will involve perceiving the status, attributes and dynamics of task-related elements in the surrounding environment. At this stage, the data are merely perceived. Though we need more perceived elements here, it is necessary for administrators to neglect others to comprehend the “Big picture” for this to work. To achieve resilience, this level includes elements of fault, security, topology, traffic, survival, and resource. Fault collects data of mis-configuration or not following correct policy which will leads to accidents [1]. Security collects alerts from different security tools such as IDS, firewall, malware detector, etc, which will result in malicious attack [31]. Changes of topology will be detected in the topology module. The onset of either a DDoS attack or a flash crowd could be initially detected by measuring traffic at the destination networks [32]. In the resource module, the missions or importance of the computers within the networks will be considered [3]. In the survival module, the deployment of survivability will be detected including fault tolerance, risk analysis and bottom-up verification [4].

B. Challenges analysis and anticipation based on information fusion

In the process of comprehension, the decision maker will form a holistic picture of the environment, comprehending the significance of objects and events. One way in accomplishing this is to build a graph from the evidence. In the resilient networks, to analyze challenge, we need to know the levels of our service, the states of our networks, state transition, and triggered condition which will be the next work. There are two steps to analyze the challenges: (1) Information extraction: we need to find, filter and categorize relevant data and extract facts and entities to understand what is happening of our networks; (2) Model generation and learning algorithms: it is necessary to
generate predefined model, train model to recognize patterns of activity. So it is necessary to define a set of resilience situation classification. Once we have defined a set of situations that we are interested in we next can store them in our DISco which can be matched with the challenge analysis.

In the stage of anticipation, we can anticipate the potential challenges which will have a possibility to lead to the transition of the network state by combing the knowledge from perception and comprehension. According to the model of comprehension, an analyst may wish to search for information that is missing from a particular model or use the model to project or anticipate where and what new data might exist.

C. Resilience estimate and analysis according to the state transition

Moitra [33] argued that system states of measuring survivability \( S \) may be \{normal, under attack, compromised, recovered, nonfunctional\}. Evaluating networks resilience in this way effectively quantifies it as a measure of service degradation in the presence of challenges to the operational state of the networks. Network can be viewed (at any layer) as consisting of two orthogonal dimensions as shown in Fig 5: one is the operational state of network, which consists of its physical infrastructure and their protocols; the second dimension is the service being provided by the network and its requirements [1]. James [34] quantified network resilience as a measure of service degradation in the presence of challenges that are perturbations to the operational state of the network. To describe the different states of the network, serial numbers are used in the set of nine \((3 \times 3)\) regions. If there is a challenge or a serial challenges happening, it will make the state transit from the best state \( S_{00} \) to the worse state \( S_{ij} \) or else. Each region may contain multiple states if the service demands such a fine granularity. If necessary, the states can be added from \( S_{00} \) to \( S_{nn} \). In the limiting case, each region represents just one state. \( S_{00} \) denotes that the system is in an initial state of “normal” for which acceptable service is delivered during normal operations. Network resilience can be evaluated in terms of the various network state transitions under the presence of challenges. When the operation is affected by challenged, which in turn may result in degradation of the service. Therefore, \( S_{22} \) denotes the state “non-functional” which means challenges make the operation severely degraded leading to unacceptable service provided in the system.

**Operational state**

<table>
<thead>
<tr>
<th>Service parameters</th>
<th>Unacceptable</th>
<th>Impaired</th>
<th>Acceptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( S_{00} )</td>
<td>( S_{11} )</td>
<td>( S_{22} )</td>
</tr>
<tr>
<td>Partially Degraded</td>
<td>( S_{10} )</td>
<td>( S_{12} )</td>
<td>( S_{20} )</td>
</tr>
<tr>
<td>Severely Degraded</td>
<td>( S_{01} )</td>
<td>( S_{02} )</td>
<td>( S_{00} )</td>
</tr>
</tbody>
</table>

Figure 5. Resilience state space [1]

To describe the state transition of the system, Petri Nets (PNs) as a directed, weighted, bipartite graph in which nodes are either places or transitions, and where arcs either lead from places to transitions (input arcs) or from transitions to places (output arcs) will be used here. PNs have ability to represent and analyze in an easy way concurrency and synchronization
phenomena, like concurrent evolutions, where various processes that evolve simultaneously are partially independent. Furthermore, PN approach can be easily combined with other techniques and theories such as object-oriented programming, fuzzy theory, neural networks. These modified PNs are widely used in computer, manufacturing, robotics, knowledge based systems, process control, as well as other kinds of engineering applications [35,36]. Garsva [37] used stochastic activity network, a flexible and highly adaptable branch of Stochastic Petri Nets, to model computer system security. Dahlbom [38] explored the applicability of a Petri net based approach for modeling and recognizing situations. In our resilient strategies, we have to deal with uncertainty. For example, an HTTP packet overflowing a buffer in the web service could be due to an application error or an attempt to gain privilege on the server. The uncertainty challenge exists in all three phase of cyber situation awareness: prior security risk management, real-time intrusion detection, and post forensics analysis [39]. Therefore, object oriented fuzzy Petri Net [35] used here models the state transition of the system. The resilience state transition model based on Petri net is illustrated in Fig.6. Each state can be described as an object, where g is a special type of transition called a gate which is a set of message passing relations among objects. In our model, g indicates a set of challenge or remediation which will result in state transition. According to the challenge and the state transition, resilience can be evaluated by validating models and measuring their effectiveness and performance. Through remediation techniques and models we can analyze whether the system can recover to a normal state. If challenges happen and meet the fire condition of the Petri model, it will lead to the transition of the network state from a better one to a worse one. Through remediation, the network state can be recovered to a better one. The possibility of state transition can be calculated which can be used to estimate the networks resilience.

V. CONCLUSION

Resilience is a dynamic and systematical property of networks which provides and maintains acceptable service level after failure happening. To make our networks resilient, it is important to “get inside” the adversary’s decision cycles (detection) to know what is going on, not just by making our own decisions quicker (remediation), but also by having better SA than the opponent (recovery). Therefore, SA is intended to combine with resilient networks in this paper. After surveying of resilient system and cyber situation awareness, resilience through SA detection model is presented according to its features, which is a part of our dynamic adaptation architecture realizing resilience control loop; then we gave the methods of challenge analysis and how to estimate resilience according to state transitions. In the next step, we will focus on challenge analysis based on SA, failure identification of fault, topology, security, traffic, survival, and resource, and resilience evaluation according to the refined failure model.

It is hoped that this exploratory paper may serve as a foundation for future research on resilience situational awareness and that resilience is recognized as critical to the design of future networks.

ACKNOWLEDGEMENTS

This research was supported by the European Union Research Framework Programme 7 via the ResumeNet project with contract number FP7-224619 and National Natural Science Foundation of China 60972078 and Gansu Province Science Foundation 1014RJZA005. The authors would like to thank members of these projects for their input into the ideas expressed in this paper.

REFERENCES


[20] Steinberg, Alan N., Christopher L. Bowman, and Franklin E. White, October 1998. Revisions to the JDL Data Fusion Model, presented at the Joint NATO/IRIS Conference, Quebec


Using Network Structure to Enhance Security and Trust of Ubiquitous Computing Services

Behnam Bazli, David Llewellyn-Jones, Madjid Merabti
School of Computing and Mathematical Sciences
Liverpool John Moores University, UK
{B.Bazli, D.Llewellyn-Jones, M.Merabti}@ljmu.ac.uk

Abstract— Developments in Ubiquitous Computing are expected to introduce interactive and configurable living spaces as an integral part of future computing environments. Such a user-oriented environment will require a secure network infrastructure to ensure integrity, interoperability, privacy, fault tolerance and simple development and execution of applications. Various middleware architectures have been proposed for Ubiquitous Computing environments as a means of securing the network infrastructure, enabling integrity, privacy and reliability as part of the service access interfaces and implementations provided. Such solutions exist as research projects, but have only been addressed and implemented within small spaces with limited functionality and services. Scalability and privacy has often been ignored by those solutions. Expanding the network infrastructure would ordinarily require the middleware to support new entities through extensive reprogramming to accommodate the change in network size. By making use of scale-free properties, the proposed infrastructure will allow the middleware to maintain a balance by rebinding to equivalent security services at specific locations within the network as the size of the network changes. This work is aimed at using small world network infrastructure and properties to propose a middleware to efficiently coordinating the components in large-scale networks.

Keywords—Ubiquitous Computing, Small World Networks; Middleware; Network Scalability; Scale Free Networks; Power-Law Distribution; Security; Privacy

I. INTRODUCTION

The interconnection of autonomous computers and isolated communication networks enables new services and applications to form distributed networks. Despite the typical centralised nature of computer networks, a distributed network operates more efficiently and effectively over a mix of workstations, LAN servers, wireless networks, regional, Web and other servers. This brings in new security challenges and the need for an appropriate security system.

Secure communication and confidentiality, tolerance to failure, ensuring availability and integrity of resources are the main concerns of distributed network security algorithms. Furthermore they maintain prevention and response to intrusion as well as functionality and interoperability of the network.

Most networks are secured by firewalls which apply packet-filtering between the internal and external network [1]. Within distributed networks the spread, mobility and locality of nodes, and the architecture of the network, makes the security measures complex. The security mechanism should consider cost, time and space as well as privacy, confidentiality, integrity and availability of resources.

Since the properties of a distributed network fit within other real networks such as human and communication networks, existing research phenomena can be applied to them, including concepts and proposals that have been brought to the field by scientists and researchers working in a variety of areas, from computing to physiology. These have all been used to model the characteristics of the nodes and entities within the network to predict future behaviour.

Work in the literature demonstrates that a number of researchers have actively studied the properties of distributed networks and their resilience to random and targeted attacks. This work has considered both how the whole network and individual nodes will be affected by such attacks [2-5].

Existing research shows that small world, scale free networks and ubiquitous computing networks share many similarities. Therefore it is imperative to study the properties of small world networks and scale free networks in order to understand the characteristics of ubiquitous computing. This will allow us to model ubiquitous computing networks and understand these systems in order to improve properties related to performance, security and usability. These networks are a growing area of research where interesting achievements and advances continue to be made.

II. DISTRIBUTED NETWORKS

A. Ubiquitous Computing

The ubiquitous computing concept was introduced because of human-centred networks. A ubiquitous computing network comprises hidden computer devices that interact with each other and humans. The layout and distribution of nodes and devices within a ubiquitous computing network, where there is no central authority to control or manage the network, makes security a complex challenge.

The deployment and use of ubiquitous computing technologies that pervade our everyday lives will have significant social and ethical impacts. It is essential to study and consider the impacts and challenges when developing ubiquitous computing systems.
To be able to tackle security concerns, we need to study and choose options that interfere least with the performance and usability while still managing to protect the network from intrusion. While the evolution of the ubiquitous computing paradigm is generally positive for users, it does introduce security concerns relating to privacy and access control, amongst other things. The security aspects of the system should therefore not be ignored. Nonetheless, security measures must be applied carefully, minimizing the impact on users and the system while at the same time minimizing the security risks present in the network. To accomplish this task we will study the structure of ubiquitous computing network with the aim of enhancing the security and trustworthiness of the network.

B. Small World and Scale Free Networks

The Small World network phenomenon was first discovered in 1967 when Milgram tried – through experimental research – to show that the average path length between two people, even when those people have only a few acquaintances, is only six steps [6]. Known as the ‘six degrees of separation’ rule, this suggests that everybody in the world is connected to everybody else either directly or using a chain of intermediary people containing at most six people.

These ideas have been examined by scientists, physiologists, marketing specialists and social researchers. This broad research has considered the properties of a variety of diverse real world networks, with the intention to model them based on their characteristics and behaviours. Small-world networks are the focus of interest because of their potential as models for interaction, economical, technological and real world networks of complex systems.

C. Clustering Coefficient

Watts and Strogatz [3] modeled the spread of a disease within a real world network. ‘Infectious diseases are predicted to spread much more easily and quickly in a small world; the alarming and less obvious point is how few short cuts are needed to make the world small’. They modeled small-world networks as a class of random graphs, noticing that these graphs could be classified using two independent structural features: clustering coefficient (CC) and shortest path length.

Roughly speaking, the Clustering Coefficient of a node in a network determines how close it is to its neighbours. Formally, for a node A, the clustering coefficient $C_A$ can be calculated as follows.

$$C_A = \frac{E}{k(k-1)}$$

where $E$ represents the number of edges that exist in the network made up of A and its neighbours, and $k$ represents the number of neighbours. Note that the maximum number of links that can occur in a network of $k$ nodes is $k(k-1)$, and hence the maximum value the clustering coefficient can take is 1, as shown in Figure 1. By contrast, for a node with $k$ neighbours, the minimum value the clustering coefficient can take is $(k-1)^{-1}$. An example of this situation is shown in Figure 2, where we can see that the clustering coefficient is lower because there are no edges among A’s first neighbours. Both the maximum and average path length increase linearly with the number of nodes and are long for pathways that have many nodes. This type of graph has been widely used as a model of an isolated signal pathway.

The nodes of a graph can be characterized by the number of edges that they have (the number of other nodes to which they are adjacent). This property is called the node degree. In directed networks we distinguish the in-degree, the number of directed edges that point toward the node, and the out-degree, the number of directed edges that start at the node. Whereas node degrees characterize individual nodes, one can define a degree distribution to quantify the diversity of the whole network’ [7].

D. Power-Law Distribution

Barabási and Albert have gone beyond the small world network and opened a new broad research field by discovering the scale-free property of real networks such as the World Wide Web, protein cells and social networks. While investigating the World Wide Web Barabási discovered that some nodes – called hubs – have many more links than others.
Further investigation revealed that distribution of nodes on the Web follows a degree distribution called the power-law distribution. A power-law distribution can be understood as occurring where a network has a large number of nodes with very low degree, tailed off to include a small number of nodes with very high degree, as shown in figure 3. High degree nodes are easier to reach and have more links than other nodes. By studying the structure of World Wide Web, Barabási discovered that unlike the theory of random distribution of the nodes most of the pages linked to just a few highly visited sites, such as Google. If a new page is created, the chances that this new site will link to Google considerably higher than for other sites that are already less frequently linked. This results in the phenomena that the ‘rich get richer’. Barabási calls this “preferential attachment.” We believe this idea can bring new prospects to information security and ubiquitous environments.

Figure 4. Power-Law Distribution.

Figure 5. Diagram of World Wide Web.

E. Preferential Attachment

Preferential attachment as a mechanism was introduced to describe how some quantities such as wealth; fame or power are distributed among individuals based on what they already have. If you search for the phrase ‘ubiquitous computing’ using Google’s search engine, 1,560,000 results are found with Mark Weiser of Xerox first on the list. The chances that a user will select this link and read his papers will be higher than for other links. The high number of reference to his papers by almost every ubiquitous computing researcher confirms this. Capital accumulation, wealth distribution and many social and scientific statistics have used a preferential attachment mechanism to model the flow of these quantities. The World Wide Web diagram created by Barabási shown in Figure 4 indicates how the Web is not the disordered network that was initially assumed.

III. Scalability

A. Security and Network Scalability

In an exponential network joining new nodes and vertices to the network without a clear structure will follow a ‘long tail’ distribution on a linear scale, as shown in Figures 3 and 5. This will have a long path that ensures a few nodes are kept away from the central hub therefore increasing the communication and response time.

Figure 6. Long tail arrangement of nodes.

The preferential attachment process of a network may lead the most popular nodes to attract connections more than others and cause overloading. Although this process keeps the average shortest path length low and the diameter of the network to an ideal and intended size, overloading hubs affects the network efficiency and service availability.

B. Privacy

The application of security to these networks is highly complex. Harold Vogt [1] demonstrated using simulations that in a small world network the end-to-end authentication of communication within the network is riskier than if applied to peer-to-peer connections between neighbouring nodes. In a distributed line of nodes $A, B, C, D, E$, as shown in Figure 6, if node $A$ wants to send a secure message to node $E$, the chance that the message will be altered is increased due to the lack of trust and the distance between them the nodes, as compared to where a direct connection exists between nodes $A$ and $E$.

Figure 7. Linear Communication.

Where a network exhibits a power law distribution the ‘preferential attachment’ may lead the network to form a long tail, with an unstructured uneven distribution of nodes which may result in the overloading of some service centres – the hubs – and therefore effect network interoperability. Without a clear structure, the connection nodes to a network may form a random network. Because of the nature of ubiquitous
computing, it’s particularly difficult to find a comprehensive solution to address scalability and privacy. However, we believe understanding the structure of such networks will be important for establishing a solution.

IV. RELATED WORK

Chen et al. [8] propose a Context Fusion Network (CFN) based operator graph model and implement this in the form of Solar, a scalable peer-to-peer platform which instantiates the operator graph at runtime on behalf of context-aware applications. Solar hosts buffer events during a disconnection period, thereby providing support for component failure and recovery. However, CFN does not support heterogeneity and privacy.

Zheng et al. [9] have used the small world network model to construct sufficient local and long range trust relationships to maintain a sparse trust between even isolated nodes. While accurate use of the clustering co-efficient and average shortest path maintains trusted relationships within the network, they have just adopted small world models without addressing privacy concern.

Al-Muhtadi et al. [10] propose a middleware for large-scale pervasive computing – an active space which is an extension of the physical space within a smart building. This highlights the scalability requirements and they propose a hierarchical approach to organise the space. However, their proposed framework fails to address the further growth of the network or introduce a tangible solution for indefinite scales, something that should not be ignored taking into account the dynamic and unpredictable growing scale of ubiquitous computing environments.

V. PROPOSED FRAMEWORK

A. Scalable Network Using Power Law Distribution

By enforcing the network to follow a power law distribution the network will become manageable and the joining nodes will be easy to organise. To some extent this will serve the network to achieve efficiency and easy development. Joining a group of hubs with associated nodes will form a network with a power law distribution and maintain the balance of the distribution. With the policies, services, resources, protocols and repositories moving to hubs this creates ‘services centre’ that maintain network reliability, interoperability and security. The preferential attachment mechanism of the network will naturally create a power-law distribution.

In a peered approach the average shortest path will stay as low as possible preventing the long tail distribution. To keep the average path as short as possible new hubs will be created to serve the newly attached links. The small world characteristics of the network will preserve the dimensions of the network naturally by keep the average shortest path between vertices low. If the number of nodes increases, more hub-service centres must be created to keep the service capability at the same level and keep the network scale manageable and interoperable.

In Figure 7 the service centres $a$, $b$ and $c$ are the existing hubs within a network. The communication between every node with each other is peered and therefore much more secure than any other approach. Each service centre will have a direct peer link of communication with the other service centres, allowing them to share any information they hold. Joining a group of new nodes to the system will create a new service centre with the same services available as the other existing hubs. We intend to derive an algorithm to limit the number of nodes attaching to each of the hubs. Therefore the efficiency will be maintained and the hub will not be overloaded with new requests. Also, to avoid disturbing the power-law distribution of the nodes, some links will be given particular roles and be allowed to attract more nodes than already defined so that demanding areas within the network remain well covered.

Figure 8. Diagram of a scalable network

B. Secure and Efficient Communication

Suppose node $B$ wants to communicate with node $C$ in Figure 7. In this case the communication will be peer-to-peer. The encrypted message will be decrypted by service centre $b$ and then passed to node $C$. The message will have an identifier so the destination node can be known, allowing the message to be routed to its hub and then onwards to the destination node.

After expansion of the network, new hubs $d$ and $e$ with associated nodes join the network. Hubs $a$, $b$, $c$, $d$ and $e$ will have direct links with each other allowing them to share the same information, directory services, protocols and policies providing access to all nodes.

Now if $A$ wants to communicate with $B$, despite having further geographical distance, the communication method and approach will be as same as for the two neighbouring nodes $B$ and $C$ as described earlier. This is possible simply because of the direct links between hubs $b$ and $d$. These two hubs will act as an intermediary to allow direct and secure communication between the two nodes. In this case the identity of the two nodes must also be known and shared by both hubs.
VI. AIMS AND OBJECTIVES

We have analysed and investigated current approaches and a number of different solutions for tackling some of the security challenges of ubiquitous computing and discovered that current tools and technologies do not support scalability of the network topology. The proposed framework intends to address and tackle the following challenges in ubiquitous computing environments.

Avoid forming random networks: communication over large area networks can be complex. Connecting new nodes and expansion of a network brings new challenges to the environment. The proposed framework retains the preferential attachment characteristic under control using an appropriate algorithm so the network does not end up forming a random network and becoming overloaded with a large number of links connected to just few hubs. We will develop an algorithm to limit the number of nodes attached to a hub. This will allow the creation of a structured and orderly arrangement of nodes so that the network maintains its existing balance and can provide services and resources to all agents. If the number of nodes linking to a hub exceeds the predefined limit, then the hub will be split into two, increasing capacity and allowing the linking of new nodes. In this way the power law distribution of the network may be disturbed and the scalability of the network effected. We allow some hubs to have more links so that they provide services to demanding areas by assigning specific roles and capabilities to some of the hubs.

Maintain short average path length: keeping the average shortest path length to a minimum will ensure trusted and direct relationships can be maintained between every pair of nodes even across long range links. The privacy of the nodes and agents will be maintained in this way using trust repetition and relationships. The average shortest path will also improve the granularity of the network. The number of nodes connected to each hub and their geographical distance can be used to improve efficiency using location sensing.

Simplify communication method: by using a peered approach and encrypting messages using private keys the communication will be much more secure than if transmitted over long ranges and through several long paths using simple encryption and decryption methods.

VII. NOVELTY

Existing middleware solutions to address ubiquitous computing network challenges do not concentrate on network scalability. One reason for this is that most of these studies have been carried out and evaluated on a small scale and in a closed environment. Of course the distribution of the nodes and unexpected and unpredictable growth of the ubiquitous computing environment means it’s important to think about the future and consider appropriate measures to overcome further challenges.

For those schemes that do try to tackle scalability, they address the issue from an engineering point of view and construct larger network by simply extending existing services and resources. The extension of a network will require secure communication methods to preserve privacy. This is the main challenge that has been ignored by almost all of the existing solutions. Therefore this work provides new possibilities for ensuring scalability and privacy within growing ubiquitous computing networks by modelling and constructing scalable network and trusted services using the small-world phenomenon.

VIII. FUTURE WORK

There is currently no comprehensive solution for addressing all of the challenges and concerns within a ubiquitous computing environment. Some of these issues, including those within the proposed framework, will be addressed in future work. The privacy and identity of users may be exposed to the system when providing a secure communication and the data will be disclosed to the hubs when encrypting and decrypting the messages. In future work we will focus on a mechanism for ensuring the privacy and identity of users remains preserved while at the same time providing a secure and peered approach for communication.

IX. CONCLUSION

The security, privacy, efficiency and effectiveness of direct communications and communication using trusted relationships has been tested, analysed and confirmed by various researchers in real world networks [1,9,11-13]. This work has inspired many information technology researchers to use similar ideas and approaches for computer networks, since both share the same small world phenomenon and properties. Therefore the mechanisms and techniques used within real networks, can be used to enhance security and provide trusted services based on trusted relationships between nodes.

The clustering coefficient and average shortest path along with the power-law distribution mechanism has been used to allocate new nodes and service centres and organise the network resources more efficiently and more effectively. A power law distribution technique is used to maintain the balance of the network and to ensure scalability as the network grows. The average shortest path length and clustering coefficient preserve direct and trust relationships between nodes, therefore keeping the trusted relationship even in a complex and growing network. Furthermore in this paper we have discussed how communication can be made secure and privacy observed and maintained using the proposed framework. We believe analyses of the results within this work will show that the approach is secure and efficient.

REFERENCES


Towards a Very High Capacity Broadband Satellite

Argyrios Kyrgiazos CCSR
University of Surrey
Guildford, Surrey
Email: A.Kyrgiazos@surrey.ac.uk

Abstract—The continuous increase in demand for broadband services indicates a requirement for high capacity satellite services to provide ubiquitous geographic coverage. As a consequence studies into the usage and applicability of adopting higher frequency bands such as Ka band Q/V bands have been undertaken. Current systems have an obvious challenge when compared to optical fibre due to the end to end delay in case of GEO satellites. Furthermore, they need to provide services at comparable cost/bit to terrestrial systems. This is also a significant challenge and requires paying special attention to the optimisation of the system as a whole. Thus, the challenge is to reduce the cost/bit and to increase the system capacity from today’s 10 GB/s to 100 GB/s or even 1 TB/s. In order to do this, a more optimum design is essential. This paper attempts to evaluate the forward link’s analysis initiated in [1], and check the feasibility of a Terabit/s satellite. Metrics such as the system’s capacity, modulations and coding statistics are calculated. Some tradeoffs are performed in order to evaluate the sensitivity of such a system and expose the characteristics that bring it closer to realisation.

Index Terms—Terabit/s satellite, satellite broadband, system architecture

I. INTRODUCTION

The telecommunications market is very competitive and demanding. Terrestrial networks have some advantages over satellite based networks and their share is constantly increasing. To compete with this fact, satellite operators are constantly striving to provide services at comparable cost per bit. Future internet demands are dramatically being increased year by year. It is foreseen that each user will demand 30 -40 Mbps up to 100 Mbps by 2020. Global IP traffic will quadruple from 2009 until 2014. The sum of video (TV, video on demand, Internet, and P2P) will continue to exceed 91% of global consumer traffic by 2014 [2]. Thus, broadband services are and will be strongly influence by the mix of traffic and in particular by video. As a matter of fact, HDTV has already reached our houses, and it has been started to give its place to 3D. A HDTV channel requires circa 7 Mbps (in DVB-S2). The 3D will demand the double of this. Recently Cisco has predicted that the overall traffic for Europe in two years will be over 104 Petabytes. The terrestrial systems are unable to cover these demands in all geographical areas and thus broadband by satellite is a key potential service provision platform. Furthermore, it is forecasted that there would be a migration from fixed speech services to mobile in the last decade, increasing the importance of satellite delivery. In addition, satellite communications have the ability to cover wide areas, where the terrestrial cannot access or provide services Another advantage is their capability to be used to good affect for low density user services for example in rural areas. Furthermore, satellites have a low carbon footprint and they do not require any street work. Current satellite systems’ capacities vary from 10 GB/s to 100 GB/s. For example, KaSAT offers a throughput of 70 Gbps, Viasat-I of 100 Gbps and Wildblue-II of 10 Gbps. Considering the traffic demands, the raw capacity will approach the Terabit/s by 2020. So the challenge is to reduce the cost/bit and to increase the capacity from today’s 10 GB/s to 100 GB/s or even at 1 TB/s. A simple way to achieve that is by widening the usable bandwidth and increasing the transmitting power and utilizing the more wisely the available resources. The usage of multi spot beams and the frequency reuse are essential to increase the spectral efficiency of a system. The main advantage of the usage of spot beams comes when the traffic is equally shared among the beams. It is well known that the type of traffic varies in space and time. Migrating to higher bands, the signal experience higher degradations, as the precipitations are more severe. For example the rain attenuation in Ka band may exceed the 20 dB. This affects services such as teleconference and telemedicine that are subject to high availability rates. Second generation satellites (Wildblue-II, Viasat-I, KaSAT) are achieving the economy of the scale due the usage of multi spot beams with high frequency reuse and adaptive modulation and coding (ACM). The question is can satellites in Ka band with current inflexible payloads support the traffic demands that arise arises and if not what can be done to fix the situation? Key technologies to answer this question are the generation of very small beams, with high accuracy and enhanced ACM for Ka band - Q/V band, providing a wider range of signal to noise ratio (SNR). But by generating very narrow spot beams, the issue of increased interference between them becomes very important. Moreover, increasing the frequency reuse factor to reduce the co-channel interference potentially reduces overall system’s capacity. In future telecommunication systems, the aspect of availability is of paramount importance. Power control and adaptive modulation and coding (ACM) techniques are not sufficient to compensate for high signal’s degradations as the ones experienced in Ka, Q/V bands. Thus, the research of new technologies is necessary. A very promising technique which combines the concept of site/antenna diversity with the utilisation of the existing infrastructure is the Smart Gateway diversity [3], [4]. It is an up-and-coming technique as the link availabilities can be increased dramatically keeping the cost to reasonable levels. In [5], the author, considering current technologies, sets the question if a Terabit system can become
a reality and examines its feasibility. In [1] and [6], the next but one generation of fixed satellite systems is analysed and the technological challenges that face this generation which is defined as operational by 2020, are presented. The authors attempt to dimension a Terabit system, identifying key issues such as constraints or essential technologies. The target of this paper is to analyse the system which is proposed in [1], and perform a sensitivity analysis. Various scenarios are examined and tested. Metrics such as total offered traffic, modulation and coding (MODCOD) statistics and C/I are extracted and evaluated. The paper is organized as follows: section II refers to the examined scenarios and the evaluation tools that were used, in section III the system’s model is presented, in section IV the results are presented and discussed, and finally in section V conclusions are presented.

II. REFERENCE SCENARIOS AND EVALUATIONS TOOLS

A. Baseline

In [1], the authors concentrate on a single geostationary satellite assuming clear sky conditions. The spectrum in Ka, Q/V bands is analysed. Proceeding to system’s architecture and having identified 4GHz and 2.5 GHz available with some constraints for Q/V and Ka band respectively, the numbers of beams that are required for the feeder and user terminal (UT) links are estimated. It is important to note that it is deemed that the optimum solution is when the feeder links operate at Q/V bands and the UT links operate in Ka band. In addition, a symmetric traffic between forward and return link was assumed. More specifically, the target is to minimize the number of beams, so a frequency reuse factor of 3 was selected. It is found out that for the UT side 88 beams with dual polarisation or 175 beams with single polarisation are needed, while for the feeder side, 55 beams are sufficient. By further exploiting the spatial separation between the gateways, and utilising the whole bandwidth in each gateway, the number of gateway beams may be reduced down to 19. It was considered that a future system will be able to operate with modulation/coding schemes that employ high spectral efficiency in clear air conditions. Thus the use of 32APSK 5/6 was considered appropriate.

B. Evaluation tools

For the purpose of the evaluation and analysis of the system, two software tools were used. The first is a Beam Plot tool, property of University of Surrey (UoS). This is a graphical tool for positioning beams and calculating the forward link’s C/I. It calculates the required antenna diameter regressively given defined beam centres, antenna pattern and beamwidth and taper roll off. For forward uplink C/I calculation, the software assumes one gateway in each beam, clear sky and equal gateway carrier power. For the forward downlink C/I, it performs Monte Carlo simulation for every beam, creating a probability distribution function of the perspective interference caused by a specific beam, and then it selects the value with the highest probability. Each point (gateway, UT) is moved on a grid within its own beam whilst interfering gateways/UT remain fixed at their respective beam centres. C/I is calculated at each possible gateway/UT location. Furthermore, it assumes that the carriers operate in TDMA, that is, one transmission per beam occupying any specific frequency. The second tool is a system simulator, provided by the German Aerospace Centre (DLR) exclusively for the SATNEX III project, and is used for the capacity’s assessment of the overall system. It is a system level simulator so it does not go into packet level or physical layer level. The main characteristics of this simulator are that it takes into account the ACM and power control. This tool is a statistical simulator that splits the overall coverage area into a grid and performing statistic computations to estimate the capacity that can be offered by the satellite to each point of the grid, and further to each beam. Effects such as rain attenuation, scintillation, polarisation tilt are taken into account [7]. Also, the system may consider non uniform traffic distribution. For the air interface, DVB-S2 is considered, but this is subject to change.

III. SYSTEM MODEL

In this section, the system model is described and any assumptions made are explained. Since the aim is to assess the analysis made in [1], the case where an operator wishes to provide broadband services via geostationary satellite across the European region was studied. For that purpose, three sets of beams patterns were generated using Beam Plot tool. The first set was the 19 gateway beams with 0.2° beamwidth, the second 88 UT beams with 0.47° beamwidth, and the last was the 175 UT beams with 0.3° beamwidth. To remind that the feeder links operate at Q/V bands, while the UT links in Ka-band. For the UT beams, an attempt for each beam pattern to cover the same area was performed. The aim was the area across Europe to be covered. In the next three figures, these sets are illustrated. In figure 1, the beam pattern of the gateway beams is presented. In figures 2, 3, the UT beam patterns, 88 and 175 beams respectively, are shown. For the UT beams, a frequency reuse factor of three was used, as suggested in [1] and for the gateway beams occupy the whole available bandwidth. Note that these beam patterns do not represent realistic implementations, as they are subject to many factors such as political or industrial issues. Their purpose is only academic and to feed the system’s analysis. In addition, the forward link air interface was considered the DVB-S2 (TDMA) and one carrier per beam. The forward user link availability was assumed equal to 99.7%. Also two types of ground terminal were considered: a typical one with 75 cm antenna’s diameter and another one representing an enterprise user with 90 cm antenna’s diameter.

From Figure 1, it can be seen that the beams are well separated from each other. Considering that the gateways reside in the centres of these beams, the distances between gateways are above 500 Km, fact which indicates very good spatial separation between them. The antenna model that is used in Beam Plot, and the one which was used to generate the necessary antenna patterns to feed the DLR system simulator is a simple antenna model that is adopted based on the illumination of a circular aperture giving results equivalent
to a single feed reflector. A tapered field distribution over the circular aperture is needed to reduce side lobes. Two taper roll-off orders were considered: $n = 1$ (Circular), $n = 2$ (Parabolic-Squared) [8]. For the reference scenarios, the taper roll-off was considered squared parabolic with an edge taper at -10 dB. The antenna’s efficiency was considered 65% as in [1]. Also, the tripoint level equals to 4.3 dB below the level at the beam’s centre.

The main advantage of utilizing an antenna with higher taper roll off is the lower side lobes levels, as it can be seen by the above figure. But in order to achieve the same beamwidth, a larger reflector is required, as the main lobe is broader. The Conventional Payload architecture is bent-pipe type payload architecture that implements the following sub-systems:

- Single Feed per Beam (SFPB) type antenna sub-system located onboard the satellite that generates a regular beam grid array of high peak directivity, narrow beam width type user beams,
- a fixed RF saturation power, TWTA based RF amplification sub-system,
- a fixed, single frequency down-conversion receiver sub-system, and
- a wave-guide based Input Multiplexer (IMUX) sub-system that performs frequency channelization. The payload only supports a fixed bandwidth to beam allocation scheme and user downlink frequency plan.

A summary of the main performance and functional characteristics of the Conventional payload is provided in TABLE I. The first refers to the case of the 88 beams, while the second to this of the 175 UT beams. As it can be easily observed, the differences between them are the reflector diameter, polarisation and the RF saturation power. For comparison reasons, it was chosen the transmitted EIRP from satellite to be the same for both scenarios. In case of 88 UT beams, the reflector antenna in satellite is smaller compared to 175 UT beams case. This results in lower directivity. Thus, a higher HPA is needed to compensate for this reduction. The antennas’ diameters were calculated automatically by Beam Plot software. As it has been already mentioned, it calculates the required antenna diameter regressively given beamwidth.

Also, for this paper, an alternating polarization type scheme in case of 175 UT beams was considered, as it is shown in the following figure.

Moreover, for the adjacent carrier filtering in transponder, a fixed value was considered equal to 20 dB. When an adjacent beam is two sub-bands away in frequency then the ratio was considered equal to 30 dB. In cases where two polarisations were utilised in the system, it was considered that the cross-polar component was 30 dB below the corresponding copolar beam peak. The polarization was taken as circular as it provides better isolation as regard as the generation of cross-polar components and easier alignment for the UT. The DLR
TABLE I
CONVENTIONAL, FORWARD LINK PAYLOAD

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antenna Sub-System</td>
<td>SFPB</td>
</tr>
<tr>
<td>Polarisiation</td>
<td>Single</td>
</tr>
<tr>
<td>Reflector Diameter (m)</td>
<td>2.82</td>
</tr>
<tr>
<td>Number of Feed Array Elements</td>
<td>88</td>
</tr>
<tr>
<td>Number of Beams</td>
<td>88</td>
</tr>
<tr>
<td>TWTA RF Saturation Power (W)</td>
<td>113</td>
</tr>
<tr>
<td>Bandwidth to Beam allocation (MHz)</td>
<td>833.33</td>
</tr>
<tr>
<td>Peak Antenna Directivity Level (dBi)</td>
<td>53.56</td>
</tr>
<tr>
<td>Typical C/I adjacent channel Isolation (dB)</td>
<td>≫20</td>
</tr>
</tbody>
</table>

Fig. 5. Alternating polarisation scheme

TABLE II
CONVENTIONAL, FORWARD LINK PAYLOAD

<table>
<thead>
<tr>
<th>Parameters / Cases</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of beams</td>
<td>88</td>
<td>175</td>
<td>175</td>
<td>175</td>
<td>175</td>
<td>175</td>
</tr>
<tr>
<td>Polarisiation</td>
<td>2</td>
<td>1</td>
<td>Alter</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Taper roll off</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Tripoint level (dB)</td>
<td>-4.3</td>
<td>-4.3</td>
<td>-4.3</td>
<td>-4.3</td>
<td>-4.3</td>
<td>-3</td>
</tr>
</tbody>
</table>

As it can be observed, the isolation between the gateways is large as the average C/I ratio is above 30 dB. Especially for the beams at the edge of the coverage area, the C/I may go up to 40 dB. With this diagram, the assumption that was made regarding the spatial separation is validated. In that way, the forward uplink does not limit the end-to-end performance. Thus, the key issue is the forward downlink. In order to provide some values to feed the DLR simulator, the C/I for the forward downlink was assessed by the Beam Plot software. In the next figure, the empirical cumulative distribution functions of C/I for the six cases are displayed.

As it can be observed, the less interference is achieved when the scheme with the alternate polarisation is used. Furthermore, the performance of case 1 is slightly worse than case 2. This is due the cross polar interference from the same beam. For the cases 4-6, it can be easily seen that their performance

IV. RESULTS

In this section, there results taken from the simulations are presented. First of all, an initial assessment for the gateways beams took place in order to support the speculation that they can be spatial separated and each gateway can occupy the whole bandwidth. The following figure shows the C/I for every point inside the beams.

Fig. 6. C/I for the gateway beams

Fig. 7. CDF of C/I for the above cases
Fig. 8. MODCOD statistics for the reference scenarios

Fig. 9. Offered Traffic

is much worse, compared to other cases. Basically, cases 4-6 represent the tradeoffs for different antenna’s characteristics. As expected, utilizing an antenna with lower side lobes and concentrating the energy to main lobe, this would result in picking less interference from other beams. Feeding the DLR simulator with results from Beam Plot, the three reference scenarios were simulated. In the next figures, the MODCOD statistics and the offered system’s capacity are shown in figure 8.

As it can be seen, the best performance is observed for the 175 beams with alternate polarisation. In this scenario, higher modulation and coding schemes can be achieved, achieving that way higher overall capacity. The worst performance applies for the 88 dual polarised beams case. Every beam suffers from the co-channel cross polar interference, which is clearly very significant. But, the differences are not so great, for example the difference between the larger and lowest value is approximately 1.3% which is translated into 6 Gbps in 477 Gbps. The main outcome of Figure 9 is that the total offered capacity is very close to 500 Gbps, which validates the assumptions in [1]. It has to be noted that the authors in [1], did not consider any atmospheric attenuation, implying that ACM can compensate for this. The MODCOD that was considered in [1] was the 32APSK 5/6. As it can be seen for Figure 7, this MODCOD is achieved by a small percentage for the case with alternate polarisation. In order to reveal the cause of this, the transmitting power for satellite for each beam was assumed 250 W, so the C/N would be so high that it has negligible effect on the link budget and the system would be exclusively interference limited. In Figure 10, the results are presented.

The target of 32APSK 5/6 is achieved and most importantly it goes up to 32APSK 9/10, achieving the highest spectral efficient scheme. This indicates that the required forward link C/(N+I) can be achieved by careful selection of satellite RF power and antenna performance.

V. CONCLUSIONS

This paper analyses and evaluates the forward link of a future satellite system. Its purpose was to justify the results in [1]. Three different cases were tested, and for different antenna’s characteristics. The main outcome of this analysis is that such a system is seemed to be feasible. The capacity of the forward link approaches the 0.5 Terabit/s. The forward link analysis indicates that the required forward link C/(N+I) can be achieved by careful selection of satellite RF power and antenna performance. By allocating the appropriate power per beam, it is possible to achieve higher MODCOD schemes, surpassing the threshold of 0.5 Terabit/s.

ACKNOWLEDGMENT

The author wish to acknowledge the contribution of ESA towards this work and for their part funding the SatNEx-III Network of Excellence under which the work was conducted. Acknowledgement is also given to DLR, Germany who manage SatNEx-III on behalf of its partners and made their simulators available exclusively for this work.

REFERENCES

