Future Work

- Control of multiple resource types (e.g. SDPs/IPs)
- Control of both IN and non-IN load
- Bid splitting taking account of SS.7 delays
- Resource availability estimation taking account of existing service sessions
- Hierarchical control
- Load control in a multi-operator environment

TELETEL
T E L E T E L

Future development of the token-based IN load control strategy will concentrate on:

- Control of multiple resource types: Depending on particular network configurations other resource types, either internal to the SCP (e.g. TCAP dialogue handlers) or external to it (e.g. Intelligent Peripheral voice circuits), rather than SCP central processors may become a bottleneck during an overload. To combat this the token generation process will be enhanced to take into account all of the resource types that will be needed throughout the course of a service session, with tokens only being allocated if access to all resource types can be guaranteed;
- The application of token-based control to the admission of IN and non-IN traffic at switches in the network will be investigated;
- A bid splitting approach based on the use of measurements of round trip delays for IN queries will be developed. In normal conditions this should have the effect of SSPs choosing SCPs that have closest physical proximity to them in the SS.7 network. It would also provide a means of dynamic re-routing of IN queries when failures in the SS.7 network causes excessive end-to-end delays. Integration of delay-based splitting with price-based splitting will also be investigated;
- Improved SCP estimation of available processing capacity, based on analysis of the future requirements of existing sessions will be developed;
- Hierarchical control, where some SSPs can bid for tokens on behalf of other SSPs will be investigated;
- The issue of load control in a multi-operator domain will be addressed.

Conclusions

- **Token-based control offers significant benefits:**
 - Effective load control during overloads
 - Maximises profit during overloads
 - Dynamic adjustment to current load conditions
 - Application-level control of service load
 - Integration of legacy SSPs possible
- **May provide longer time-scale load management:**
 - Adjustment to daily/weekly load patterns
 - Dynamic re-configuration of routing when resources are introduced or withdrawn
- **However: Significant changes to IN equipment!**

TELETEL
T E L E T E L

This presentation has argued the case for the benefits of adopting a network-oriented approach to load control in the rapidly evolving Intelligent Network domain. Network-oriented load controls provide a means of co-ordinating the decisions regarding the admission of service requests across an entire IN, thereby affecting network-wide protection from overloads in a manner that can dynamically adjust to changes in traffic patterns. Use of well-designed network-oriented controls mean that the operator is given a greater guarantee that the network will never overload; therefore they need not be as over-dimensioned as currently deployed ones. This would lead to significant cost savings for the operator.

A network-oriented load control strategy based on the use of tokens to control admission of service requests was described and its performance analysed. The results presented showed that the token-based strategy performed significantly better than traditional, node-oriented controls, its main advantages being that it provides effective protection against resource overloads, it is maximises generated profit, it can provide application level differentiation between service classes, it facilitates dynamic routing of IN queries between multiple SCPs and allows for the integration of legacy SSPs.

In this presentation the token-based strategy was described as a very responsive control, which tracks sudden changes in traffic patterns or resource availability. However it could operate on a longer time-scale, with token allocations taking place less often. For example token generation could be used to "re-configure" the network on an hourly basis to take account of the changes in the usage patterns of services over the course of a day. On an even longer scale token generation could be used to re-configure the network upon the introduction or withdrawal of resources. In the case where a specialised SCP, optimised for a particular service class is introduced into the network price-based bid splitting would ensure that this SCP is used (in preference to other SCPs) by SSPs for requests of that service class.

Evaluation of scalability in a broadband mobile communication system

Georgios Karagiannis (**)
Victor Nicola (*)
Ignas Niemegeers (*)

(*) University of Twente
(**) University of Twente and Ericsson

0-7803-6317-5/00/\$10.00 ©2000 IEEE

Contents

- Introduction
- Objectives
- Network architecture
- Performance measures and models
- Scalability studies
 - Increasing the number of users per node
 - Increasing the number of nodes
- Conclusions

♦ The use of mobile communications has grown rapidly during the last two decades. Broadband mobile communication systems, such as the broadband version of the Universal Mobile Telecommunication System (UMTS), is expected to provide an increasing number of fixed and mobile advanced services to a rapidly growing number of users. Therefore, it is crucial to consider scalability issues relating to the operation of these networks in commercial and public environments.

♦ In this paper we investigate the scalability of a broadband mobile communication system such as a broadband version of the future UMTS signalling system, that is considered in the European ACTS project EXODUS [ACTS95]. This broadband UMTS version uses the Intelligent Network (IN) concept to provide a flexible introduction of new advanced fixed and mobile services. We call the broadband IN infrastructure as Broadband IN (B-IN).

♦ By using simulation experiments, we study the impact of processor speed (in physical entities) and network configuration on the signalling performance.

This presentation is organised as follows:

- ♦ First, we describe the concepts used.
- ♦ The objectives of this presentation are explained.
- ♦ Network architecture and topology used in the scalability experiments are described.
- ♦ The performance models (i.e., user workload models and network models) and measures used for the performance evaluation are introduced.
- ♦ The experiments and performance results are discussed.
- ♦ Finally, the conclusions.

Introduction

- Broadband Intelligent Network (B-IN):
 - separated service control and service switching
 - different B-IN core network architectures:
 - specified by International Telecommunication Union (ITU)
 - IN CS1, IN CS2, IN CS3
 - narrowband scheme specified by Internet Engineering Task Force (IETF)
 - RFC2719: "Framework Architecture for Signaling Transport"

- ♦ The Intelligent Network (IN) concept provides an architectural separation between service control and service switching aiming at supporting flexible introduction of new advanced services, e.g., video telephony.
- ♦ The future generation of IN should be able to support a mix of broadband fixed and mobile services. Such an IN, that we will call Broadband-IN (B-IN), has been defined in the European ACTS projects INSIGNIA and EXODUS [ACTS95].
- ♦ The B-IN core or backbone network could be either the Broadband Integrated Services Digital Network (B-ISDN) or the Internet Protocol network.
- ♦ The B-ISDN type is specified by the International Telecommunication Union (ITU) standardisation body, wherein three IN Capability Sets (CS1, CS2 and CS3) are defined (see [Vet98]). The CS3 should specify the introduction of broadband fixed and mobile services. The communication network layer is the Asynchronous Transfer Mode (ATM).
- ♦ The Internet Protocol type is currently studied in the Internet Engineering Task Force (IETF) standardisation body. An IN architecture that is supporting narrow-band services and is using the Internet Protocol layer as its communication network layer, has been specified in the Request For Comments document RFC 2719 [RFC2719].

Introduction

- ♦ The figure at the left depicts the B-IN architecture that is using a B-ISDN (ATM) core network, while the figure at the right is using an Internet Protocol core network.
- ♦ The main physical entities constituting such a basic B-IN architecture are:
 - ♦ Broadband Service Switching Point (B-SSP)
 - ♦ Broadband Service Control Point (B-SCP)
 - ♦ Radio Access System (RAS) that it consists of:
 - ♦ Cell Site Switch (CSS)
 - ♦ Base Transceiver Station (BTS)
 - ♦ Mobile Service Control Point (M-SCP)
- ♦ Broadband Intelligent Peripheral (B-IP) which provides the specialised resources that are required for the provision of IN broadband services
 - ♦ Fixed Terminal (FT) represents the end fixed user
 - ♦ Mobile Terminal (MT) represents the end mobile user
- ♦ The Signalling System 7 (SS7) may use one of the following as a communication network layer:
 - ♦ ATM layer
 - ♦ Internet Protocol layer
- ♦ In this work we focus on the B-IN architecture with B-ISDN (ATM) core network. However, we expect that the conclusions are also valid when the Internet Protocol network is used as a core network.

Scalability

- **Scalability:** ability to increase “size” of network while maintaining “network performance”
- Network “size”:
 - number of users supported by a network node
 - number of network node and links
 - geographical spread covered by the network
 - number of services provided by the network
 - size and distribution of accessible B-SCP data base

♦ Scalability (see, e.g., [Gau96]) is one of the most important factors in the design of a distributed multimedia system, such as B-IN. Network scalability can be defined [KaBe97], [Ka198a], [VeHu98] as the ability to increase the “size” of the network, in some sense, while maintaining QoS and network performance criteria.

♦ The “size” of the network may relate to one of the following:

- ♦ *the number of users that must be supported by a network node*
- ♦ *the number of network nodes and links*
- ♦ *the geographical spread covered by the network*
- ♦ *the number of services provided by the network*
- ♦ *the size of the data objects*
- ♦ *the size and distribution of the accessible B-SCP database*

Scalability

- How to evaluate Scalability?

- creating and maintaining a balanced network:
 - (1) load balancing to provide equal load on B-IN physical entities with similar functionalities, e.g., B-SSPs
 - (2) configure B-IN (e.g., the CPU processing and/or memory access speeds) to obtain equal utilizations for the B-IN physical entities with different functionalities, e.g., B-SSPs and B-SCPs
- increase the ‘size’ of the network and measure performance
- State-of-the-art work on IN load balancing, using a distributed communication platform
 - CORBA: Ericsson’s Jambala platform, achieves linear scalability
 - JAVA: JAIN platform
 - Mobile Agents: European ACTS project MARINE

♦ The scalability is evaluated for a balanced network (which is necessary for the efficient use of network resources).

♦ Load balancing may be achieved by using distributed communication platforms. Some examples are:

- ♦ the Ericsson’s Jambala [Jambala] platform is using the CORBA (Common Object Request Broker Architecture) distributed platform and is supporting a mix of mobile and fixed services. Jambala can provide linear scalability by allowing the network operators to expand their network nodes, e.g., number of processors, to accommodate an increasing number of users.
- ♦ the JAIN platform [JAIN] is using the JAVA platform to provide a framework that builds and integrates networks and services among various networks, e.g., ATM, Internet Protocol. This platform can increase scalability by integrating new nodes and/or networks into an existing configuration.
- ♦ the European ACTS project MARINE [MARINE] that uses the results from other projects such as INSIGNIA and EXODUS and introduces mobile agent technology to increase the scalability by providing flexible allocation of the service control resources.

B-IN Services

- IN fixed service
- B-VPN (Broadband Virtual Private Network):
 - Call handling
- IN Mobile service
 - Mobile Telephony:
 - Call handling
 - Location Update
 - Handover

Objectives

- present performance models for scalability studies of B-IN
- propose a method to evaluate the scalability of B-IN
- experiments to emphasize the need for scalable broadband mobile communication networks

- ♦ The B-VPN service [Insig101] is an IN fixed service that realises a logical sub-network of a B-ISDN which appears to a specific group of users as a private broadband network, for voice, video or data communication. This service incorporates call handling procedures, i.e., call setup and release.
- ♦ The Mobile Telephony service is an IN mobile service, that incorporates call handling (call setup and release), location update and handover procedures:
 - ♦ Location update is used when a mobile user (terminal) reports its location to the B-IN network upon entering a different Location Area (LA).
 - ♦ The handover procedure can be described as the mechanism that automatically transfers a connection (call in progress) associated with a roaming MT from one BTS supporting a cell to another BTS supporting a neighbouring cell.

- ♦ In this work we investigate the scalability of a broadband mobile communication system such as a broadband version of the future UMTS signalling system, that is considered in the European ACTS project EXODUS.
- ♦ We present performance models and a method to evaluate the scalability of a B-IN.
- ♦ By using simulation experiments, we emphasise the need for scalable broadband mobile communication networks.

Network Architecture

- Network Topology: One B-IN island

- ♦ This slide depicts the topology used in the first set of experiments, where the ability of a B-IN island to support an increasing number of users is investigated. It consists of:

- ♦ four B-SSPs
- ♦ one B-SCP
- ♦ B-IP is not included, since it is not required for the B-VPN or Mobile Telephony services that are considered here.
- ♦ Each B-SSP is inter-connected with two Radio Access Networks (RANs). A RAN consists of:
 - ♦ one CSS,
 - ♦ one or more BTS,
 - ♦ one M-SCP
 - ♦ a varying number of MTs.

- ♦ Since we focus on the impact of mobility procedures on the signalling performance of the B-IN core and intelligence network parts, we take into account the access delays due to the RAN without considering its detailed internal functionality, i.e., we view the RAN as a set of end mobile users.

Network Architecture

- Network Topology: The B-IN network

- ♦ This slide depicts the topology used in the second set of experiments, where the network scalability is evaluated for an increasing number of B-IN islands.

Signalling Information Flows

- Specified in ACTS/INSIGNIA and ACTS/EXODUS for:
 - Broadband Virtual Private Network (B-VPN)
 - Call Setup/Release
- Mobile Telephony service
 - Call Setup/Release
 - Location Update
 - Handover

- ♦ The signalling information flows describe the interactions required among the physical entities to establish an IN service request.
- ♦ The B-VPN signalling information flows (Message Sequence Charts: MSC) for call setup and call release are defined in the ACTS INSIGNIA project [Insig1201].
- ♦ The Mobile Telephony call handling, location update and handover message sequence charts are specified in the ACTS Exodus project [Exod1103], [Exod1206].

Performance Measures

- Network throughput for B-VPN and Mobile Telephony
- Mean setup delays for B-VPN and Mobile Telephony
- Mean delays of Location Update and Handover procedures

- ♦ The performance measures are the throughputs and mean delays of:

- ♦ B-VPN call setup
- ♦ Mobile Telephony:
 - ♦ call setup
 - ♦ location update
 - ♦ handover

- ♦ The B-VPN (resp. Mobile Telephony) setup delay is defined as the time duration from the instant when the user (resp. RAN) starts the B-VPN (resp. Mobile Telephony) call setup procedure until the call setup procedure is completed, excluding any user (resp. RAN) response times.

- ♦ The location update delay is defined as the time duration from the instant a RAN starts the location update procedure until the location update procedure is completed, excluding any RAN response times.

- ♦ The handover delay is defined as the time duration from the instant a RAN starts the handover procedure until the handover procedure is completed, excluding any RAN response times.

User Workload Model

- Poisson request arrival process:
 - 33.33 % B-VPN (Video telephony) calls
 - 66.66% Mobile Telephony calls:
 - 75.5% Location Update
 - 21% Call Setup/Release
 - 3.5% Handover
- Call duration pdf is a mixture of two normal pdf's on a logarithmic time scale:

$$f(t) = a \bullet f_1(t) + (1 - a) \bullet f_2(t), \quad -\infty < t < \infty$$

- ♦ The user workload model describes the patterns and intensities of the call setup, the location update and the handover requests supported by the B-IN network.
- ♦ For the Mobile Telephony service it is assumed that the calling and the called terminals are mobile or fixed with equal probabilities.
- ♦ Furthermore, for both B-VPN and Mobile Telephony services, the called user is connected to any B-SSP in the B-IN network with equal probability.
- ♦ Assuming that the total traffic (due to mobile and fixed services) per user is fixed, the load per network node can be varied by varying the number of users per node.

Network Model

- Each B-IN physical entity : a single server queue
- Each network island : a queuing network model
- Message routing : B-VPN and Mobile Telephony MSCs
- Time-Based priority scheduling:
 - higher priority to messages belonging to earlier user requests
 - handover requests assigned higher priority than all other requests

- ♦ The network models represent the flow and processing of messages through the network on behalf of user call requests and mobility procedures.
- ♦ Each B-IN network island is modelled as a queuing network (see next slide). The details of routing and processing times of messages through this queuing model are based on performance measurement results accomplished in the INSIGNIA project (for preliminary results see [Kan98b]).
- ♦ In our performance experiments, the so called *Time-Based* priority scheduling discipline is used (see [Insig1106]). Using this priority scheduling, new and cycled messages belonging to the same B-IN request (i.e., a B-VPN call setup or a mobility procedure) get the same (non-pre-emptive) priority, which is assigned based on the arrival time of the B-IN request. Earlier B-IN requests and their messages are assigned higher priorities than those assigned to later B-IN requests and their messages.
- ♦ To insure a seamless handover, this procedure is assigned a higher (non-pre-emptive) priority than all other B-IN requests and the associated procedures.

Network Model

- Queuing network model of an B-IN island

- ♦ Each physical entity in the B-IN core and intelligence parts (i.e., B-SSP and B-SCP) is modelled by a single server queue with an infinite buffer.
- ♦ The RANs and fixed users are modelled as infinite server queues.
- ♦ The total network with many B-IN islands could be represented in a similar way by extending the queuing network model depicted in this slide.

Scalability Studies (Increasing the processing speed)

- B-VPN and Mobile Telephony throughputs at utilisation 0.8

- ♦ For a balanced network, the throughputs of all procedures (i.e. B-VPN and Mobile Telephony call setup, location update, and handover) increase linearly with the processing speed at the physical entities (i.e., B-SSPs and B-SCPs).
- ♦ The 95% confidence interval of all simulation results are within $\pm 4\%$ of the corresponding estimates.

Scalability Studies (Increasing the processing speed)

- B-VPN and Mobile Telephony mean delays at utilisation 0.8

- ♦ The mean delays of all procedures decrease with the processing speed.

Scalability Studies (Increasing the number of B-IN islands)

- B-SCP database access delay depends on the number of B-IN islands (N):
 - constant: B-SCP database access delay is independent of N
 - logarithmic: B-SCP database access delay is proportional to the natural logarithm of N
 - linear: B-SCP database access delay is linearly proportional to N

- ♦ It is assumed that the total load on the B-IN network is equally divided among all islands.
- ♦ The experiments were performed by increasing the number of B-IN islands N , from 1 to 8.
- ♦ Let $d(N)$ be the B-SCP database access delay when the number of B-IN islands is N . We consider three types of dependency:
 - ♦ Constant: $d(N) = a, N \geq 1$, where a is the B-SCP database access delay.
 - ♦ Logarithmic: $d(N) = a + b \cdot \ln(N), N \geq 1$, where b is a constant that is set to $1/3$.
 - ♦ Linear: $d(N) = a' + b' \cdot N, N \geq 1$, where $a' = a - b$.

Scalability Studies (Increasing the number of B-IN islands)

- Mobile Telephony call handling throughputs at utilisation 0.8

- ♦ When the B-SCP database access delay is a constant independent of N , the throughputs increase linearly with the number of B-IN islands.
- ♦ When the B-SCP database access delay is proportional to N , the throughputs of all procedures exhibit a sub-linear (i.e., less than linear) relation with the number of B-IN islands. This is due to the fact that as N increases, the B-SCP database access delay increases, thus decreasing the throughput of the B-SCP.
- ♦ As expected, the throughputs for a logarithmic dependency falls between those for the constant and linear dependencies.

Scalability Studies (Increasing the number of B-IN islands)

- Mobile Telephony call handling mean delays at utilisation 0.8

- ♦ For an increasing number of B-IN islands (N):
 - ♦ Mean delays of all procedures either remain almost constant (when the B-SCP database access delay is independent of N) or
 - ♦ decrease slightly (for linear and logarithmic dependencies on N).

Conclusions

- Present performance models for scalability studies of B-IN
- Provide a method for evaluating the scalability of broadband mobile communication networks (e.g., B-IN)
- Emphasize the need for scalable broadband mobile communication networks (e.g., B-IN)

Conclusions

- The method is demonstrated by simulation experiments
 - initially, the network is balanced
 - increasing processing speed while fixing the network utilisation:
 - Throughputs increase linearly
 - Mean delays decrease

Conclusions

- increasing the number of B-IN islands (N) while fixing the network utilisation and introducing B-SCP access delay dependency on N :
 - For constant dependency: throughputs increase linearly with N
 - For logarithmic (resp. linear) dependency: throughputs exhibit slower (resp. much slower) than linear increase with N
 - Mean delays almost unchanged
- Current work:
 - Study of load balancing and congestion control issues for a B-IN that supports an increasing number of advanced fixed and mobile services

References:

- [ACTS95] ACTS'95, "Project Summaries", 1995.
- [Exod1103] "Service Description and Network Architecture for phase II", reference project number AC013/ITL/WP1.1/DS/P/003/b1, May 1996.
- [Exod1206] "Interfaces and Protocols for Phase II", reference project number AC013/EXODUS/WP1.2/DS/P/006/c1, October 1996.
- [Gau96] Gauthier E., Le Boudec J.-Y., (1996) Scalability Enhancements on Connection Oriented Networks, in *Lecture Notes in computer science*, Vol. 1044, pp. 27-38, 1996.
- [Insig1101] "First Trial: Description of the Selected Services", reference project number AC068/CLT/11/DS/P/001/c0, July 1996.
- [Insig1106] "Network-based Signalling Traffic Models and Performance Parameters", reference project number AC068/UOT/114/DS/P/006/b0, August 1997.
- [Insig1201] "First Trial: Protocol Specification", reference project number AC068/NTU/120/DSL/001/x1, 1997.
- [JAIN] de Keijzer J., Tait D., Goedman R., "JAIN: A New Approach to Services in Communication Networks", *IEEE Communications Magazine*, January 2000.
- [Jambala] Jones F., "Jambala - Intelligence beyond digital wireless", *Eriasson Review* N0. 3, 1998.
- [KaBe97] Karagiannis G., Beijnum van B.J., and Niemegeers I.G.M.M. "On the integration of the UMTS and B-ISDN system", Seventh IFIP Conference on High Performance Networking (HPN/97), 1997.
- [KaNi98a] Karagiannis G., Nicola V.F., Niemegeers I.G.M.M., "Quantitative Evaluation of Scalability in Broadband Intelligent Networks", *International Conference on Performance of Information and Communication Systems (PICS'98)*, pp. 65-82, May 1998.
- [KaNi98b] Karagiannis G., Nicola V.F., Niemegeers I.G.M.M. "Performance Impact of Integrating Mobile Services in Broadband Intelligent Networks", 7th IEEE Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks, CAMAD'98, August 1998.
- [MARINE] ACTS project AC340/MARINE, "Mobile Agent Environments in Intelligent Networks", information located at: <http://www.uk.infowin.org/ACTS/RUS/PROJECTS/ac340.htm>
- [RFC2719] Ong L., Rytina I., Garcia M., et. al., "Framework Architecture for Signalling Transport", IETF RFC 2719, 1999.
- [Veltu98] Venieris I., Hussmann H., (editors) "Intelligent Broadband Networks", John Wiley & Sons, 1998.