The Mobile Application Part (MAP) of GSM

JAN A AUDESTAD

Jan A. Audestad is Senior Adviser in Telenor

This is the story of how the Mobile Application Part was developed. MAP is the neural system interconnecting the distributed computer infrastructure of GSM (VLRs, MSCs, HLRs and other entities). The work on MAP started as a study of the general architecture of mobile systems in ITU two years before the GSM group was established. In 1985 the work was taken over and later completed by the GSM group. MAP was the first protocol of its kind in telecommunications systems. After long and difficult negotiations with the community of switching and signalling experts, MAP got its final structure in 1988. This included the use of Signalling System No. 7 as carrier and the support of services such as roaming, call handling, non-interruptive handover, remote switching management and management of security functions. MAP is one of the real technological triumphs of GSM.

Interconnecting operators: The three sausage model

This is the story of how the Mobile Application Part (MAP) was created and the problems and conflicts that this development caused between the communities of enthusiastic mobile communications experts and the conservative and preservative telephone switching engineers.

The idea to design an international land mobile communications system came up in CCITT[1)] Study Group XI (Switching) in late 1980, more than two years before GSM was established. The first document describing how mobility was achieved in the NMT system was presented to the Plenary Assembly of CCITT in 1981. The document was appended to a proposal for a new study question on land mobile systems. The authors were Bjørn Løken and myself. The Study Group found the proposal so interesting that they appointed Bjørn Løken as Interim Rapporteur so that work on land mobile systems could commence immediately without awaiting the approval of the Plenary Assembly.

From mid 1981 the work on land mobile systems in CCITT took off at a violent speed and important results were obtained early. The three most important results obtained during the early years were a simple network model nicknamed the "tree sausage" model, a method for non-interrupt handover of calls between different switching centres, and the outline of a protocol supporting mobility within and between Public Land Mobile Networks (PLMN). The latter was later christened the Mobile Application Part (MAP). These achievements were all adopted and developed further by the GSM group.

Figure 1 shows the "three sausage" model of a land mobile system. The sausages are the two PLMNs and the fixed network. The significance of this model is that it is entirely abstract; that is, independent of the particular physical design of the network. The model can be used to describe all major activities taking place in the system and it is general enough to apply to every practical configuration of a mobile system. This allowed MAP to be developed independently of a particular network architecture: MAP is thus not specific for GSM.

The PLMN represents an entity of network ownership. The interaction between PLMNs is thus a cooperation of different mobile network operators allowing mobile subscribers to roam between networks independently of ownership and subscription.

The model shows two PLMNs together with the fixed network. MAP supports mobility between the PLMNs, that is, MAP offers to mobile terminals the capability to move from one PLMN to another PLMN retaining the capability to receive and make calls. Interconnectivity between a mobile terminal and a fixed terminal or between two mobile terminals takes place across the fixed network. The signalling protocol connecting the PLMN to the fixed network is Signalling System Number 7 (SS No 7). The plan was to implement SS No 7 in the telephone network during the 1980s so when an international mobile system was ready for operation around 1990, this would be the natural choice of interconnection method. When GSM was put in operation in 1991/92, SS No 7 was in place and mobile communications could commence immediately.

1) *The Consultative Committee on International Telephony and Telegraphy of the International Telecommunications Union (ITU).*

PLMN architecture: The heritage of NMT

Another development that took place prior to GSM was the specification of an internal architecture of land mobile systems as shown in Figure 2. The architecture is based on the principles first developed for NMT and were adjusted for use in more flexible configurations by the CCITT.

The architecture is simple. In addition to the radio infrastructure (base stations and cells) and the telephone exchanges switching the calls between the base stations and the fixed network (MSC), there are two types of databases: VLR (Visitor Location Register) and HLR (Home Location Register)[2]. The VLRs are in charge of a location area. Within this area the mobile stations can roam without updating their location[3]. Location updating takes place when the mobile station roams from one location area to another. The VLR contains all information about the mobile terminals currently in its location area required for establishing calls to and from these terminals. The VLR also controls the switching process in the MSC. In this respect, the VLR is an intelligent network (IN) node similar to the concept developed independently by Bellcore during the 1980s for serving telephone calls requiring centralised control (routing and payment management of free-phone and premium rate calls, and management of distributed queues and helpdesk functions). In mobile networks, identity management, location updating, routing administration and handling of supplementary services require the support of similar procedures as intelligent networks. Therefore, it is not surprising that the two groups came up with rather similar solutions to the problem of remote control of switching processes independently of each other.

The HLR is a database containing subscription information (services and capabilities) and the current location of the mobile terminal. There is usually one HLR for each GSM operator[4].

In addition, there are also other databases and entities not important for the basic architecture (equipment registers, authentication key escrows, voicemail systems and short message centres). These databases are also connected to the other network elements by MAP.

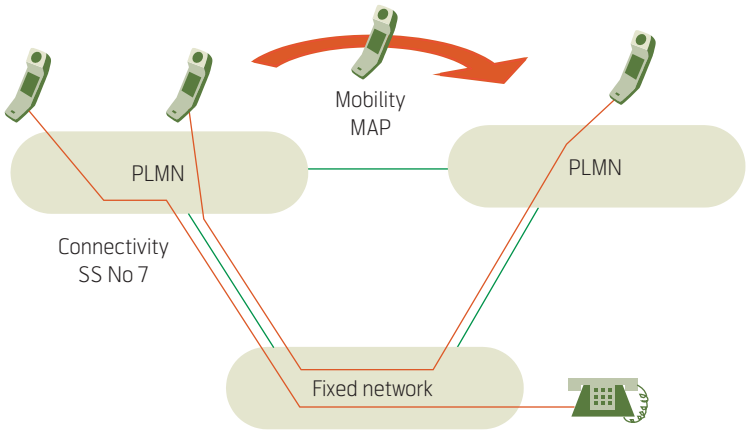

Figure 1 The three sausage model

There are permanent MAP connections between the VLR and the MSCs it is controlling. There are sporadic MAP connections between the VLR and the HLR for location updating, between two or three MSCs for handover, and between two VLRs for management of identities during location updating. Sporadic means that a relationship is established only when two such entities must exchange information and controls.

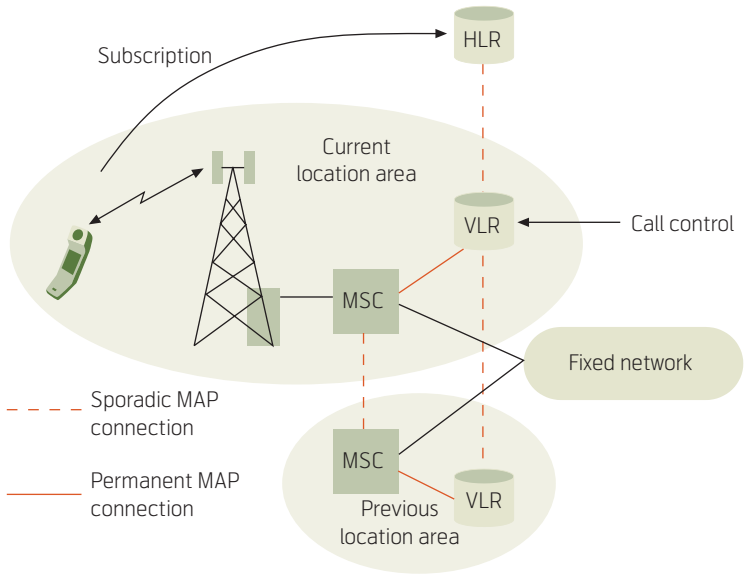

Figure 2 Architecture of GSM

---

2) *These names were inspired by NMT: HMX for home mobile exchange and VMX for visiting mobile exchange offering similar functions as the HLR and the VLR, respectively.*

3) *A VLR may control a number of location areas. Roaming between such location areas only require updating of the VLR and not the HLR.*

4) *If there are many mobile subscribers, two or more HLR databases may be required. Such configurations are regarded as a single, distributed HLR.*

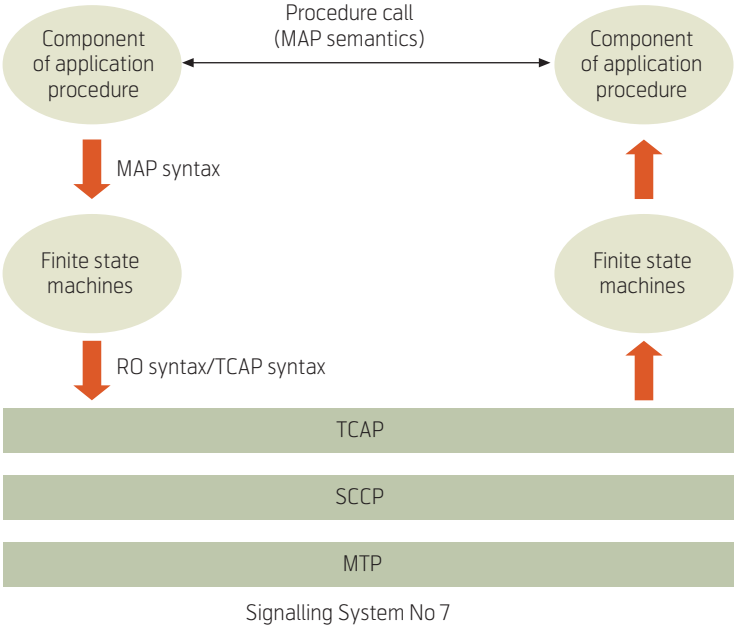

Figure 3 MAP protocol

Many of the operations and procedures supported by MAP were outlined in CCITT before the GSM group was established. The work had in fact progressed far before GSM took over the specification in 1985.

Cooperation between SPS/SIG and GSM

Until 1985, not many of us believed that Europe would ever be able to specify a common land mobile system. CEPT[5)] had this far never succeeded in its standardisation efforts so there was no reason to believe that it would succeed this time either. At the GSM meeting in Berlin in October that year, this attitude changed. The core team of the GSM group had now become tightly consolidated and it was decided unanimously that every effort should be made to make GSM a success. An outline of 110 recommendations that should comprise the complete GSM specification was drafted at the meeting so that work could commence on solving concrete problems. Amongst these recommendations was the MAP specification.

This was the kick-off the GSM group needed.

CEPT already had a working group dealing with signalling, SPS/SIG. This group was asked to take charge of the MAP specification. Christian Vernhes, who had worked on MAP right from the beginning in CCITT, became chairman of the team in charge of the specification. In addition, the team consisted of Bruno Chatras (ASN.1 specification) and myself (procedures). Alfred Karner joined the team in the autumn of 1988 cleaning up the formal aspects of the specification.

The request from the GSM group was to deliver a complete specification by January 1989. SPS/SIG met this requirement. Both the syntax and the semantics of the protocol had then been methodically tested and verified by Siemens. This included testing the ASN.1 syntax of the protocol and all procedures that were written in SDL.

Protocol structure

The final structure of the protocol is shown in Figure 3. GSM is a distributed system where most of the procedures are divided into components located in different computers, for example, in the mobile terminal, in one or two VLRs and in the HLR for location updating. A handover between cells connected to different MSCs involves simultaneous execution of software components in the mobile terminal, in two base station tranceivers, two base station controllers, two or more MSCs and one VLR. The components in the MSCs and the VLR are interconnected by MAP. The other components are interconnected by other protocols.

MAP is the core element in the network architecture of GSM, tying together MSCs, VLRs, HLRs and other specialised databases and equipment. The configuration of the protocol is shown in Figure 3.

The MAP semantics corresponds to the exchange of procedure calls between the components of the distributed application procedure as shown in the figure. The syntax and the semantics of these procedure calls are mapped onto the protocol via a set of finite state machines in order to transport the commands across the network. The machines not only map the syntax and the semantics but also take care of software synchronisation, ensure that the right software components are bound to each other during the whole transaction, and detect and respond to exception conditions. The state machines ensure that the abstract syntax of MAP can be interfaced to any application layer protocol. MAP can in fact be interfaced with any application layer protocol supporting procedure calls, for example RPC (Remote Procedure Call) of the Internet.

5) *Conférence Européenne des Postes et des Télécommunications. The part of CEPT in charge of telecommunications standardisation later became ETSI (European Telecommunications Standards Institute).*

The protocol stack on which MAP was finally implemented is shown in Figure 3. The stack consists entirely of sub-protocols of SS No 7: TCAP (Transaction Capabilities Application Part), SCCP (Signalling Connection Control Part), and MTP (Message Transfer Part).

In search of a suitable protocol stack

The reliability of MAP must be as good as in telecommunications networks at large. This means that the maximum outage of the protocol should not exceed 15 minutes per year which is the standard reliability requirement for telephone exchanges, telecommunications links and signalling systems. This corresponds to the ultrahigh availability of 99.997 % – that is, the system must function properly 99.997 % of the time. Therefore, the stringent requirement on availability puts constraints on the choice of protocol stack for supporting MAP: all layers in the stack must be at least as good as this.

Another important requirement is the time budget as shown in Figure 4. The time budget has to do with psychology. The time from a telephone call is initiated and until the called user is connected (i.e. activation of the ringing tone) must not be more than one or two seconds; otherwise the calling user may prematurely clear the call because a few seconds is a very long time when waiting for the ringing tone. The total time must be divided among a large number of processing events allowing 100 milliseconds or less for each of these processes. In the GSM system, such an event may be information exchange between the VLR and the HLR during call establishment. On average the exchange of messages, including processing time at each end of the connection, should not require more than 100 milliseconds on average and less than 200 milliseconds for 90 % of the calls. Broadly speaking, this allows for typically 30 milliseconds one-way delay (5000 km) and 20 milliseconds processing time. The delay includes processing time at any intermediate node such as router for directing the signalling information.

The availability requirement and the timing constraint put severe restrictions on the choice of protocol stack.

Only three protocol stacks suitable for transfer of MAP messages were available in the mid 1980s. These are the Internet suite of protocols, the protocols of public packet switched data networks (X.25), and Signalling System No 7. One problem was that these stacks were not even complete; that is, they did not contain all layers that were required for supporting the application protocol.

The most complete protocol stack was RPC over TCP/IP. MAP fitted nicely into the RPC format, TCP ensured end-to-end integrity and IP was able to route the messages across the network. Around 1985, the Internet was only used as a research network. The network was owned by no-one and there was no formal organisation in charge of its evolution. The reliability and quality of service of the network was unspecified except that messages were delivered on a best effort basis, that is, there were no specification concerning how much time it would take to transfer a message across the network. The Internet as such was therefore not suitable as a carrier of information requiring real time operation and ultra-high availability.

Nevertheless, one possibility we had was to apply the Internet *technology* between MSCs, VLRs and HLRs – not the Internet itself. Exploitation of the Internet technology offered us one possible, though not the wanted, solution. We kept this as the last resort solution all the time up to the summer of 1988.

Another possibility was to use the X.25 data network. A complex stack of application protocols, the Open System Interconnection (OSI) suite, was specified by CCITT and ISO during the 1980s. The problem with the X.25 network was that it did not contain secure enough mechanisms for autonomously restarting operation after a major link failure. These had to be built into the specification before X.25 could be used for high reliability transfer of information. Furthermore, the network supporting MAP had to be a dedicated network for MAP only because otherwise other traffic could cause freeze-out or excessive delays. X.25 did not offer more than what the Internet technology did for a fraction of the cost. X.25 was therefore never a serious alternative.

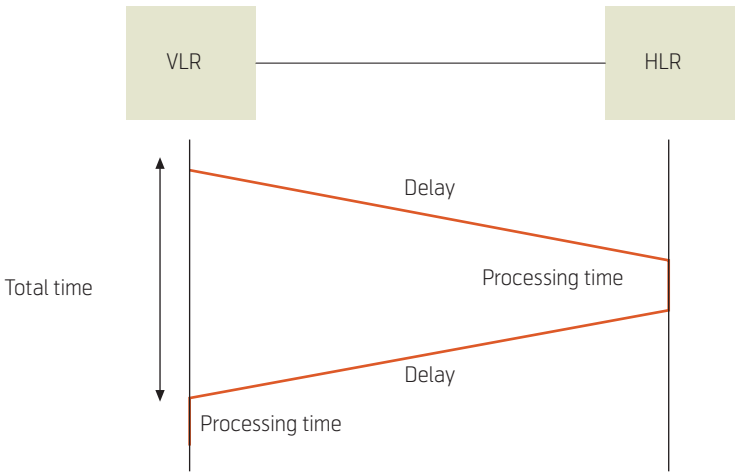

Figure 4 Time constraints

The solution we wanted was Signalling System No 7 because this protocol was designed for extreme reliability and real time operation. However, the signalling system did not support higher layer protocols ensuring end-to-end integrity. The basic protocol designed for operation between telephone exchanges consisted only of a data link layer and a simple network layer that could only be used on a very simple network infrastructure. This is the Message Transfer Part (MTP). In order to convert the signalling system into a data network supporting flexible routing, the Signalling Connection Control Part (SCCP) was developed. This is essentially a data network combining the features of IP and X.25, offering both connection-oriented and connectionless operation. However, the SCCP had a separate connectionless category offering guarantee of delivery. The signalling system offered neither a transport protocol nor application layer protocols. We needed these protocol layers in order to ensure proper end-to-end control of the messages.

We requested the signalling subgroup of Study Group XI already in 1984 to specify such protocol layers but our request was ignored. It came as a surprise when we received a question from the same group in 1986 asking us to provide them with the set of requirements we needed to implement MAP. The group had started specifying an application layer protocol called the Transaction Capabilities Applications Part (TCAP). We learned much later that the reason this was done was not to satisfy our needs but to support the development of the intelligent network in the USA.

This time things also went wrong because Study Group XI specified a protocol that was adequate for the simple intelligent network protocols but was far too simple for a complex protocol like MAP. We had to make some additions to the specification before it could be used.

Transaction Capabilities (TCAP): The best we got

TCAP consists of two layers: the component sublayer supporting the Remote Operations (RO) protocol and the transaction sub-layer protocol. RO is a remote procedure call protocol standardised as part of the OSI development. It allows two application entities to exchange invoke commands requesting that a remote task is executed, return of the results of the execution and provision of error codes if things go wrong.

The transaction layer is a connection-oriented protocol where the information generated by RO is sent in Begin, Continue and End messages supporting obvious functions. The protocol offers multiplexing since several RO messages can be sent in one transaction message.

However, it turned out that TCAP was too simple for a protocol as complex as MAP. TCAP did not offer appropriate mechanisms for controlled establishment and release of the connection, explicit binding of the application modules, activity management, recovery management and software synchronisation. It never occurred to Study Group XI that there were good reasons for the complex application layer protocols defined by ISO for general data communications: association control, session management, concurrency management and reliable transfer of information.

The functions needed for proper operation of the GSM system (association control, explicit binding, session management, and recovery) were included in the MAP specification itself, making it more complex and less universal than necessary.

Alignment with the fixed network: Much hindsight, little foresight

One of the major problems we faced was the alignment of GSM and the fixed network. The general opinion of the fixed network community, including the market analysts of the telecommunications operators, was that mobile communications would remain a small service as compared to fixed telephony in all future. Statistics from the Nordic countries showing a steady and tremendous growth of mobile communications was brushed aside with the argument that this was just a transient that would soon be over.

This attitude was global, where the strongest opponents were the North-American telecommunications operators for which mobile communications represented a real competitive threat. Mobile communications was a way in which new companies could perforate the monopolies of the seven RBOCs[6)] – nicknamed the Baby Bells – that AT&T was sliced into during the divestiture of AT&T in 1984.

The small but growing community of people working on mobile systems had come to the opposite conclusion: the design of handheld mobile terminals had become feasible, moving the mobile phone out of the car and onto the pavement. The potential market for mobile communications was then just as big as the

6) *Regional Bell Operating Company.*

market for fixed telephony, or perhaps even bigger: telecommunications was about to enter a new era where flexibility and ubiquity would become the most important drivers of market growth.

One undisputable requirement that we had to accept right from the beginning was that the GSM system should not have any impact on the specifications of the fixed network. This included signalling, routing algorithms, numbering and number analysis, switching functions and network architecture.

We may regret this attitude now.

Let us look at some problems that this decision caused not just for the GSM system but for telecommunications in general.

GSM uses non-geographic numbers. This means that there is no relationship whatsoever between the telephone number of a mobile station and where that station is located. The number only identifies where in the network the subscription information and the current location of the subscriber can be found. We had to invent new methods by which we could handle the routing of calls to mobile terminals since this was the first time that the problem of non-geographic routing occurred in real systems.

For several years the MAP team and the GSM group proposed that the routing algorithms of the telephone network should be altered allowing a look-ahead procedure in MAP by which the location of the called mobile subscriber was first found and then the call was sent directly to the destination. This procedure was proposed in order to avoid “tromboning”; that is, to avoid that a call from an Italian to a Norwegian on vacation in Italy is set up all the way from Italy to Norway and then back again to Italy. The application of a look-ahead procedure would ensure that the call could be established on a short connection within Italy. At this time, the look-ahead procedure was regarded as a feature solely required in GSM. However, if implemented, the look-ahead procedure would have solved several routing problems such as number portability, allocation of personal numbers, role and time dependent routing of calls, and selection of the shortest and cheapest route for voicemail calls, free-phone calls, calls to distributed helpdesks, calls to premium rate services, and so on. It would also have enabled a complete integration of GSM and intelligent network services.

The feature was neither accepted by the CCITT nor by ETSI. This decision resulted in many problems later as the use of non-geographic numbers has become the rule rather than the exception.

Another problem was concerned with segmentation of messages in the SCCP. No such service was included in the original specification of the SCCP since the opinion was that segmentation of long messages had to be done by the application using the SCCP. It took us several years and numerous meetings to persuade the designers of the SCCP that segmentation at the application layer is impossible because there is no simple way in which the length of application messages can be determined. The reason is that the application protocol is written in an abstract syntax requiring compilation in exactly the same way as any program written on the computer. The compilation of the messages takes place at the interfaced between TCAP and the SCCP and it is only there that the exact bit pattern of the message is known. The length of a message then depends on both the syntax of the MAP message and the format of TCAP: the bit pattern is different for the same MAP message if it is sent in a Begin or a Continue format. The length and content of the bit pattern depend on such a subtle detail as the invoke number of the RO message and the transaction number of the transaction message. Segmentation must then either take place in TCAP after compilation or in the SCCP before the message is sent. The message must be reassembled before it is delivered to the de-compilation process of TCAP wrapping up the transaction messages and the remote operations in order to pick out the MAP message.

The task to make the signalling experts understand this subtle detail was difficult but in the end the SCCP was amended in order to support segmentation. Otherwise we had been forced to resort to solutions based on TCP/IP.

A final problem was concerned with addressing. SCCP offered three addressing schemes: the point codes identifying exchanges, subsystem number intended to identify a specific box or function at the signalling interface, and the global title which is nothing but a telephone number identifying the termination. The point code is not globally unique and cannot be used for MAP; MAP required so many subsystem numbers in order to identify the different functions and boxes that too many of the available codes were used up. We could therefore only use the global title. This seems trivial but it is not.

In data communications, several addresses are usually needed in order to identify the process that needs to be invoked. The global title identifies only the VLR, HLR, MSC or whatever other type of equipment is found at the termination. The global title does not identify each of the several functions that the VLR is in charge of. Noting that the VLR may not be a single

computer but a complex of many computers makes the problem even bigger. This means that the final destination of a TCAP message – the application module that shall perform the actual task – can only be identified from addressing information contained in the transaction messages, the remote operations messages or the MAP messages.

In order to understand that this is a problem, let us look at how this feature is solved in TCP/IP.

Broadly speaking, the IP number identifies the computer. The header of IP contains a field indicating which type of protocol is contained in the information field of the packet, e.g. TCP or UDP. This allows the computer to select the correct software for analysing the content of the information field. Similarly, TCP (or UDP) contains an address called the port number. The port number identifies a computer program that is capable of analysing the application message embedded in the information field of the TCP datagram. Behind each port there is then a particular software that is able to interpret the format of the message (http, email, mpeg and so on). Based on the analysis of addresses and other information included in the application message the computer can then finally start the program that can actually handle the content of the message.

When the SCCP and TCAP were specified, this way of handling computer interactions was obviously not understood properly. Otherwise, a transport layer would have been introduced just in order to handle addresses beyond simple routing addresses in complex computer systems. A protocol as simple as UDP would have done the trick.

Without the support of a transport layer protocol and supporting application layer protocols, MAP relies too much on Signalling System No 7 and became less flexible than we hoped when we started developing it.

In the end a success

MAP was the first application protocol designed for signalling purposes. The first version of it was sent for approval of the GSM group in February 1989. The work on the second version had already started a couple of months earlier. In less than one year the second version was finalised, in time for implementation in the first GSM network. Version 1 was never implemented. The first version contained more than 650 pages including appendices and supplementary material; the second and third versions contain about 750 and 1000 pages, respectively.

The first version consisted of 54 individual procedures. The newer versions are just a little bigger.

It took much time to develop the first version of MAP. The reason was that there was no similar protocol that we could use as template. In fact, we had to make all the mistakes ourselves in order to learn how to do it. The ASN.1 encoding of the protocol, for instance, was rewritten several times before we understood how to write it. I think we misinterpreted the ASN.1 language in all the ways this possibly could be done. Finally, Alfred Karner showed us how to do it.

The application protocols developed later for intelligent networks, supplementary services in the telephone network, universal personal telecommunications and UMTS are all based on the methods and principles developed in MAP. Some of the members of the MAP team were invited to various groups in CCITT and ETSI to explain how to interpret the ASN.1 and SDL specifications and apply these languages in protocol design.

The development of the MAP protocol was indeed pioneering work.

For a presentation of the author, turn to page 54.