GIG/DISN Quality of Service and Service Level Agreement Management for Integrated Global Wireless Tactical Services Provider Network

Dr. Syed Shah and Bruce Bennett
Defense Information Systems Agency
Arlington, VA

and

Pamela Hemmings
Booz Allen Hamilton
Herndon, VA

ABSTRACT

As the Defense Information Systems Network (DISN) transitions to a native-IP network supporting converged unclassified and secret data, voice, and video on the Global Information Grid (GIG) backbone, an extensive Quality of Service (QoS) architecture will be required to effectively support real-time and mission critical traffic. Service Level Agreements (SLAs), which define the negotiated and contracted service between the Defense Information Systems Agency (DISA) and it's customers (DoD Services and Agencies), will rely on QoS policy implementations to ensure contracted service levels can be satisfied. Effectively managing SLAs can be challenging in fixed, wireline environments given the dynamic nature of packet-based traffic and varying mission priorities and requirements of the DoD. Extending these services into wireless tactical environments with mobile users and varying wireless link conditions and network topologies introduces additional complexity.

With the emergence of standards-based Commercial-off-the-Shelf (COTS) technologies including 802.16, 802.20, and other OFDM-based technologies, broadband wireless networks may soon provide a last-mile tactical extension of the DISN/GIG. Supporting real-time and mission critical services across these wireless networks involves traditional IP QoS mechanisms as well as additional link layer QoS mechanisms to dynamically and intelligently allocate RF spectrum among multiple users. To provide a more seamless extension of the DISN/GIG, these wireless networks must be capable of maintaining QoS and SLAs that adhere to the GIG's End-to-End QoS policy. This paper addresses the following topics: the deployment of a GIG End-to-End QoS policy that meets the needs of the disadvantaged tactical warfighter; the challenge of deploying and managing a consistent End-to-End QoS policy when using network hardware with differing QoS capabilities; and the evaluation of QoS and SLAs Management, and Policy-Based QoS.

INTRODUCTION

The GIG/DISN is the DoD's premier global, end-to-end information transfer infrastructure. DISN provides a robust communications infrastructure and services needed to satisfy national defense, command, control, computing, communications, and intelligence requirements and meets corporate defense requirements. The DISN includes the Unclassified but Sensitive Internet Protocol (IP) Router Network (NIPRNet); the Secret Internet Protocol (IP) Router Network (SIPRNet); the Defense Red Switch Network (DRSN); the Defense Switch Network (DSN); the DISN Video Services (DVS); and TRANSPORT Services. All the separate networks are converging into one integrated network and the current DISN services shall transition to native-IP services provided on the GIG backbone. Significant cost reduction may be possible due to the more efficient utilization of network resources through packet switching, as well as the need to maintain only one network infrastructure [1].

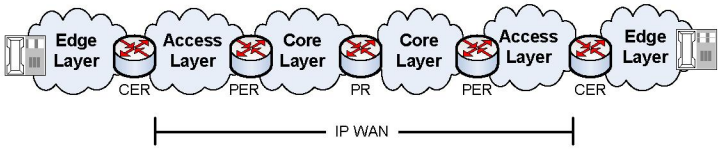
Convergence is not only occurring between circuit-switched and packet switched networks, but also between mobile and fixed networks. In this paper, we will introduce an architectural design of the QoS management and SLA concept on end-systems, specifically concentrating on the access layer at the tactical edge. The integration of QoS technology management and SLAs support will provide enhanced means for more effectively supporting the warfighter's dynamic transport requirements in a policy-based management environment. In the end, we will also discuss some open technical issues to support end-to-end QoS for quality-aware, multi-media services in a mobile, heterogeneous, and multi-domain environment consisting of terrestrial, IP SATCOM, and broadband wireless networks.

GIG/DISN ARCHITECTURE

For purposes of this paper, the GIG network is divided into three layers[1]: Edge, Access, and Core. DISA's role is to install, operate, maintain and manage this totally integrated, interoperable, protected, flexible and reliable global telecomm infrastructure. The Edge Layer is associated with the Local Area Network (LAN) and the Campus Area Network (CAN). The boundary for the Edge Layer is the customer edge router (CER). The Edge Layer is considered robust and the LAN/CAN characteristics include high bandwidth, diversity, and redundancy. Edge

Layer Quality of Service (QoS) is predominately provided by means of the high bandwidth. The Access Layer connects the Edge Layer network to the Core Layer via a MAN or circuit that connects the CER to a Provider Edge Router (PER). The Access Layer may or may not include limited bandwidth and diversity. In addition, the Access Layer may consist of a meshed network, fiber optic ring (i.e., MAN), or point-to-point circuits.

The Core Layer is composed of a high speed optical network and contains two types of DISN Service Delivery Nodes (SDNs). The first type of SDN is defined as robust and is characterized by high bandwidth, diversity, and redundancy. The second type of SDN is not considered robust due to its lack of bandwidth and it may also lack diversity. A router that connects two PERs is classified as a Provider Router (PR). The WAN is defined as the portion of the network consisting of the Access and Core Layers. Figure1 shows the GIG/DISN architecture with different network layers.



Legend:

PR – Provider Router (Core Router)
PER – Provider Edge Retailer (Access Router)
CER – Customer Edge Router

Figure 1: GIG/DISN Architecture

GIG/DISN QOS MECHANISMS

Each DISN service can be differentiated by the SLA which defines parameters such as the QoS and the required bandwidth. SLAs have no real value in themselves. Their value lies in the way in which they are managed in the network. It is essential to improve DISA's (service provider) ability to meet the contract with the customer, as defined by the SLA [2], in order to make optimal use of the network while minimizing any penalties from non-compliance. In this section basic QoS mechanisms for IP networks are outlined. It is inevitable that congestion will occur in parts of the network and that some services, due to their strict requirements, should be given differential treatment by introducing classification and per hop treatment policies as well as bandwidth reservation mechanisms. Scalability and cost of such solutions to assure QoS are key factors when evaluating the following QoS technologies and architectures: Differentiated Services (DiffServ), Integrated Services (IntServ), and MPLS (Multi-Protocol Label Switching) Traffic Engineering.

Integrated Services: IntServ leverages admission control signaling protocols such as RSVP to reserve network resources end-to-end before traffic is transmitted and to notify users of the availability of those resources [3]. Implementing an IntServ-based solution can introduce scalability and complexity issues in terms of the signaling overhead as well as the management of thousands of user traffic throws. The inflexibility and complexity of IntServ can be amplified in highly dynamic networks. To improve the scalability of IntServ, an aggregate resource reservation solution can be used in the core to reduce the number of flows and provide a more scalable approach.

Differential Services: DiffServ-based packet forwarding involves a combination of classifying, marking, metering, and shaping packets at the network edges and scheduling and queuing packets in the core nodes. DiffServ uses a field in the IP header, called the DiffServ field, as the DiffServ Code Point (DSCP) to classify packets from different traffic flows. Packets are differentially forwarded on a per-hop basis according to their DSCP and the policies configured on each router or switch. This per hop treatment offers a scalable approach but does not alone provide sufficient guarantees for high priority users[4,5].

MPLS Traffic Engineering: MPLS provides a more efficient, faster method of providing QoS using Layer 2 forwarding. With MPLS, each packet is assigned a forwarding equivalency class (FEC) only once, as the packet enters the network. The FEC is encoded as a label and is sent along with the packet. At subsequent nodes, the network header of the packet does not need to be re-examined; the MPLS label is used as an index to a table that specifies the next hop and the next label. MPLS will enable traffic engineering which selects traffic paths in order to optimize network utilization and meet traffic requirements [6,7].

A combination of these QoS mechanisms will be implemented in different parts of the GIG/DISN to create an effective yet flexible and scalable QoS architecture. The QoS architecture will be rolled out in a phased approach, initially supporting somewhat simpler mechanisms, increasing in capability and dynamicity with each phase.

GIG/DISN QOS AND SLA MANAGEMENT

The true success of QoS mechanisms and architectures lies in the effective enforcement of QoS policies to satisfy SLAs. With the transition to a converged IP network, the configuration and management of QoS mechanisms and rules will become an extremely dynamic task. A Policy-Based Network Management (PBNM) solution will be

required to dynamically translate mission priorities and requirements into QoS policies and enforce those policies. PBNM solutions take in user-defined, high-level policies, translate them into device-level configurations and commands, and automatically implement them at the network element level. The PBNM solution dynamically applies the policy across all network devices through a unified interface. A PBNM solution consists of policy servers that provide storage, decision making, distribution, and policy monitoring services. Policy agents run on the network elements and enforce policies. Policy Decision Points (PDPs) or policy servers make decisions based on policy rules and the state of services that those policies manage. Policy Enforcement Points (PEPs) or agents run on the device or network resource and enforce the policy decision and/or implements configuration changes. Figure 2 illustrates a logical PBNM hierarchy.

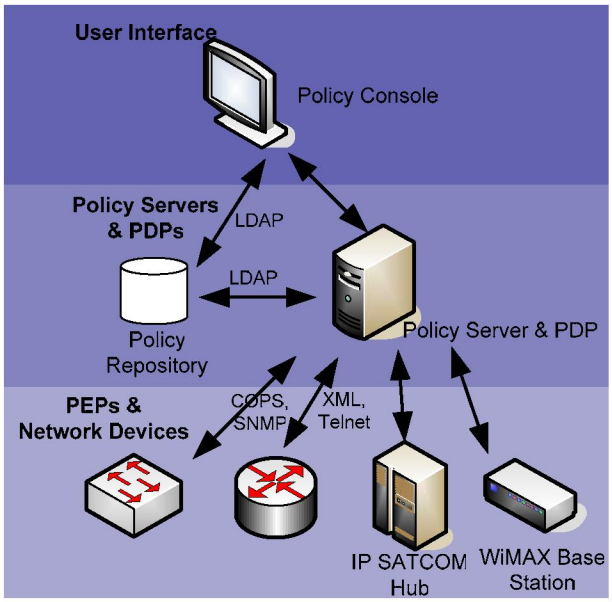


Figure 2. Logical PBNM Diagram

As illustrated in the diagram, it is essential that the policy server communicate with all network devices including traditional switches and routers as well as IP SATCOM hubs and wireless base stations which control access to the network. Providing consistent, end-to-end QoS will require dynamic, automatic configuration of all networking devices and must support consistent QoS capabilities across a heterogeneous transport system, consisting of terrestrial, SATCOM, and wireless links.

To support SLA management, the PBNM solution will be integrated with the SLA provisioning system to ensure network resources can satisfy new and existing SLAs before services are provisioned. The following sections describe the DISN/GIG as a tactical services provider network.

TACTICAL SERVICE PROVIDER NETWORK

With the recent developments of the DoD Joint IP MODEM effort, the presence of IP SATCOM terminals in-theatre providing two-way connectivity back to the terrestrial GIG/DISN will become increasingly prevalent. Emerging broadband wireless technologies, such as WiMAX, are being examined to further extend broadband services from fixed SATCOM terminals out to mobile users with small form-factor devices. Figure 3 illustrates the network architecture proposed in the Tactical Service Provider (TSP) Advanced Concept Technology Demonstration (ACTD).

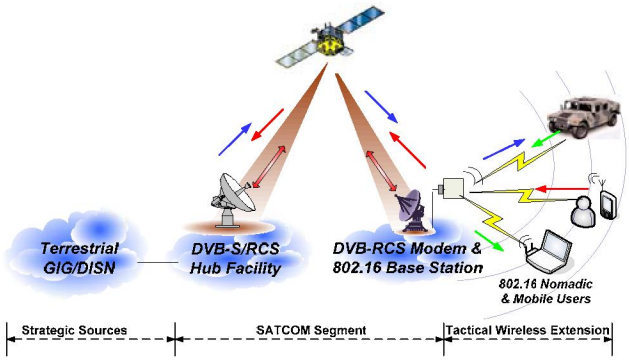


Figure 3. Proposed TSP ACTD Architecture

In the proposed architecture, the terrestrial GIG/DISN is extended into theatre using a two-way IP SATCOM System consisting of Digital Video Broadcast – Satellite, Next Generation (DVB-S2) on the forward link and Digital Video Broadcast – Return Channel Satellite (DVB-RCS) on the return link. DVB-RCS uses Multi-Frequency Time Division Multiple Access (MF-TDMA) to efficiently allocate return link bandwidth among multiple end user modems. The DVB-S2/RCS modem connects to the 802.16/WiMAX base station to wirelessly extend connectivity to mobile users. Mobile WiMAX uses an Orthogonal Frequency Division Multiple Access (OFDMA) technology to efficiently allocate bandwidth among multiple 802.16 mobile subscriber units. Effectively supporting real-time and mission critical services over this tactical service provider network will require QoS mechanisms on the SATCOM and wireless portions of the network.

IP SATCOM RESOURCE ALLOCATION & QOS CAPABILITIES

As the work of the Joint DoD IP MODEM work develops, IP SATCOM technology, specifically DVB-RCS, will

become a common part of the GIG/DISN's access layer network. The value of using DVB-RCS technology to support IP traffic over a SATCOM system lies in the efficiency of MF-TDMA technology. With traditional SCPC (Single Channel Per Carrier) SATCOM, terminals are defined with a static frequency and bandwidth. Channel bandwidth is allocated at all times while the terminal is logged on whether or not packets are being transmitted. Conversely, MF-TDMA allows a group of users or Satellite Interactive Terminals (SITs) to communicate with a gateway using a set of frequencies, each of which is subdivided divided onto time-slots. The Gateway will allocate to each user (SIT) a series of time slots, each defined by a frequency, bandwidth, start time and duration. In MF-TDMA, bandwidth utilization is reduced to only what is needed at a given instant in time. This will free up capacity that would have been wasted in an SCPC allocation. "Free" bandwidth can then be utilized by another new user (SIT) or by an exiting user to increase its throughput.

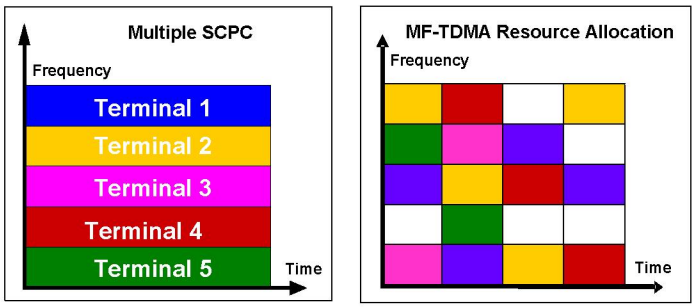


Figure 4. MF-TDMA vs Multiple SCPC Service

As IP SATCOM terminals become common elements in the DoD's network, special consideration must be made to fully understand and take advantage of the system's resource allocation and QoS capabilities. In terms of managing return link bandwidth, the DVB-RCS standard defines several forms of capacity assignment: Continuous Rate Assignment (CRA), Rate Based Dynamic Capacity (RBDC), Volume Based Dynamic Capacity (VBDC), Absolute Volume Based Dynamic Capacity (AVBDC), and Free Capacity Assignment (FCA). The capacity requests and assignments are typically made on a frame-by-frame basis. CRA provides a committed information rate capability without the need for constant bandwidth requests and is intended for services requiring a constant bandwidth such as VoIP without silence suppression. FCA allocates remaining bandwidth after all other bandwidth requests have been satisfied to eliminate wastes in empty time slots. Typically, the hub is statically configured with minimum and maximum allocation values for each end user modem. When the hub receives bandwidth requests from all of the modems for a given timeframe, the hub will analyze the total available bandwidth along with each modem's request and their configured min and max throughput values to determine how bandwidth will be allocated for the next frame. This process enables granular prioritization or QoS among individual modems.

Table 1. DVB-RCS Resource Allocation descriptions [10]

Capacity Allocation Type	Description	Potential Services
CRA Continuous Rate Assignment	Rate capacity provided for each and every frame	VoIP
RBDC Rate Based Dynamic Capacity	Rate requested dynamically, overrides previous requests, & can include time-out value	Streaming video or audio
VBDC Volume Based Dynamic Capacity	Volume requested dynamically & is cumulative with previous requests	FTP, data transfer, HTTP
AVBDC Absolute Volume Based Dynamic Capacity	Volume requested dynamically & replaces previous requests	Used when VBDC requests may have been lost
FCA Free Capacity Assignment	Rate capacity assigned which would be otherwise unused	Not assigned as it is extra free capacity

To support finer levels of QoS that differentiate according to service, IP address, protocol, UDP or TCP port, or DSCP value, vendors must implement additional traffic classifiers. The mapping of IP DiffServ classes, with associated classification, conditioning and scheduling functions, into MAC layer classes and their associated DVB-RCS capacity categories is required to ensure IP QoS requirements are accordingly enforced at the layer 2 access layer. Vendors are currently in the stage of implementing this capability, concentrating first on VoIP support. Some variation of this DiffServ support is typically implemented by vendor solutions. In addition, vendors are investigating support for IntServ and resource reservation signaling protocols. Full support for this entire suite of QoS capabilities will be required for seamless integration of an IP SATCOM system with the terrestrial GIG/DISN.

IEEE 802.16 BANDWIDTH ALLOCATION & QOS CAPABILITIES

To effectively extend real-time and mission critical broadband services across mobile, wireless networks, additional consideration must be given to the inherent bandwidth allocation and QoS capabilities of the equipment. With an OFDMA system, the base station assigns each mobile subscriber station a number of transmission burst slots, defined by time and frequency, during which the subscriber may transmit on the uplink. The base station broadcasts an uplink burst map on the downlink which assigns the time and frequency slots for all of the subscriber stations. The number of slots assigned to an individual subscriber station directly correlates to that subscriber's uplink bandwidth. WiMAX applies fast scheduling in both downlink and uplink where the scheduling can change very quickly and the amount of resources allocated can range from the smallest unit to the entire frame. This is especially well suited for bursty data traffic and rapidly changing channel conditions.

Accordingly, the base station decides how bandwidth is allocated among subscribers according to their current bandwidth needs, prioritization, and SLA. The QoS section of the 802.16 standard defines a number of bandwidth request and scheduling mechanisms in order to differentiate service levels for a number of services. To provide point-to-multipoint, shared access between multiple subscriber units, the WiMAX base station must efficiently allocate spectrum among multiple users in order to satisfy QoS requirements on a per subscriber basis as well as a per service basis. The IEEE 802.16 standards define a connection-oriented MAC layer. To support the mapping of services to subscriber station and their associated varying levels of QoS, all data communications are in the context of a transport connection. A transport connection defines both the mapping between peer MAC layers (at the subscriber and base station) and a service flow. The service flow defines the QoS parameters for the data exchanged on the connection. Service flows provide a mechanism for uplink and downlink QoS management and are integral to the bandwidth allocation process.

The QoS parameters defined in the service flow will typically include minimum reserved traffic rate, maximum reserved traffic rate, traffic priority, request/transmission policy, as well as scheduling type. The standard defines several scheduling options, which correspond to the service classes described in the Table 2.

Leveraging a variety of bandwidth request and scheduling mechanisms, the 802.16 QoS architecture supports: differentiated levels of QoS - coarse-grained (per user/terminal) and/or fine-grained (per service flow per user/terminal), admission control, and bandwidth management. WiMAX systems will have the ability to map DSCPs and MPLS flow labels to the 802.16 bandwidth allocation mechanism described below [9].

Table 2. 802.16 Bandwidth Allocation Types and Corresponding Applications [9]

Bandwidth Allocation	Applications	QoS Characteristics
UGS Unsolicited Grant Service	VoIP	• Max. Sustained Rate • Max. Latency Tolerance • Jitter Tolerance
rtPS Real-Time Packet Service	Streaming Audio or Video	• Min. Reserved Rate • Max. Sustained Rate • Max. Latency Tolerance • Traffic Priority
ErtPS Extended Real-Time Packet Service	Voice with Activity Detection (VoIP)	• Min. Reserved Rate • Max. Sustained Rate • Max. Latency • Tolerance • Jitter Tolerance • Traffic Priority
nrtPS Non-Real-Time Packet Service	File Transfer Protocol (FTP)	• Min. Reserved Rate • Max. Sustained Rate • Traffic Priority
BE Best-Effort Service	Data Transfer, Web Browsing, etc.	• Max. Sustained Rate • Traffic Priority

Systems will also support the implementation of policies as defined by various operators for QoS-based on their SLAs (including policy enforcement per user and user group as well as factors such as location, time of day, etc.) [9].

While the IEEE has defined an extensive QoS architecture for the 802.16 standard, initial compliance and interoperability testing conducted by the WiMAX Forum

addressed basic interoperability at the RF and MAC layers but did not include QoS profiles. As a result, early versions of certified equipment may have limited QoS functionality. However, a few vendors have implemented fairly extensive QoS capabilities and provide an interface at the base station or access controller to configure QoS parameters. For example, the base station or access controller will allow network operators to assign prioritization levels according to individual subscriber stations, application type, source and destination addresses, as well as defined classes of service (CoS). Currently, the provisioning of service and ability to assign priorities and QoS levels resides in either an individual base station management system or an access controller system that controls an entire network of base stations. While the configuration is somewhat static and manual in current systems, integrating a PBNM system with the base station and access control systems will enable a much more dynamic and capable QoS architecture.

PERFORMANCE CONSIDERATIONS IN A WIRELESS ENVIRONMENT

Providing QoS over a mobile wireless network compared to a fixed terrestrial network introduces a variety of challenging issues. Whether supporting fixed or mobile services, link conditions on a wireless network can vary greatly from one location to another depending on terrain (flat, hilly, or mountainous), building clutter (dense urban, suburban, or rural), building types and materials, as well as other sources of interference (other operators or subscribers, self-interference, household device interference, etc.).

Introducing mobility creates a very dynamic network with varying link conditions depending on each user's location at a given time. For example, a user may have a very strong signal in a line-of-site location at a given time and be capable of transmitting at high data rates. Then the mobile user may slip behind a building, experience significantly lower signal strength, and accordingly only support much lower data rates. To support these varying link conditions, the 802.16 standard defines adaptive modulation which addresses how the subscriber unit and base station automatically adjust modulation levels and data rates to account for varying link conditions. Under good link conditions, the equipment can afford a less robust modulation level and coding rate with increased data rates. However, under poor link conditions, the equipment will use a more robust modulation level with additional coding at the cost of decreased data rates.

The variation in link conditions and available user data rates further validates the necessity of providing a dynamic resource allocation mechanism in the 802.16 base stations. On a frame-by-frame basis, base stations dynamically modify uplink and downlink scheduling to adapt to changing link conditions and bursty traffic conditions of each subscriber unit and eliminate resource inefficiencies due to empty time slots.

INTEGRATION OF HYBRID NETWORKING QOS CAPABILITIES FOR END-TO-END QOS

As previously described, the GIG/DISN QoS architecture will consist of a variety of QoS mechanisms in different parts of the network to achieve an effective yet manageable QoS implementation. While different mechanisms will likely be implemented in the respective core, access, and edge layer networks, these individual implementations must be integrated to support an end-to-end level of QoS. Significant effort must be made to provide similar, consistent QoS in SATCOM and wireless transport systems. These systems must have the ability to translate IP and MPLS QoS markings into their respective service classes or service flows and coordinate resource scheduling mechanisms appropriately. To achieve a more seamless QoS implementation across this heterogeneous network, a number of guidelines have been described below.

First, the QoS parameterization shall be independent of the actual QoS solutions used at lower levels within the network, and of the transport technologies used in the network. The QoS signaling shall convey appropriate QoS information to describe the QoS requirements of the IP flow, session or connection. But it may also be appropriate to convey QoS related information to describe the current network QoS condition along the bearer path. The QoS information negotiated with the backbone network shall be stored in the Policy Server.

In the GIG/DISN, all the routers, switches and other network access devices must have a sophisticated queuing strategy that will process the packet according to the different QoS requirements. Implementing a set of queuing strategies on an ultra-high-speed router/switch that can provide a wide variety of QoS guarantees is not trivial, and has yet to be fully realized in today's high-speed routers/switches. Second, to guarantee QoS requires the cooperation of all routers/switches along the transit path of the packet. If one router along the transit path cannot guarantee the QoS for the packet, all the guarantees from other routers are wasted. This aspect makes the full support of QoS especially difficult when the packet traverses multiple domains with different administrations. For these reasons, the support of fine granular QoS in the IP network has not been fully materialized. Most service

providers today offer the simplest service level agreement (SLA) to their customers—the average transit delay between various nodes within the network [10].

CONCLUSION

This paper provided an overview of IP QoS Architectures, the relevant standardization work and points out areas where further study is needed. Considering the current activities of DoD/DISA on GIG/DISN networks, it is fundamental to have a systematic understanding and background for making recommendations on traffic handling and QoS. Increasingly, the end-to-end and interdomain treatment of traffic will be important in an all-IP scenario. In this field much work remains to be completed. When new net-centric services are deployed in GIG/DISN DoD networks, and with the aim of introducing an all-IP platform, it will be necessary to take into account the end-to-end QoS requirements of such services. This is a complex problem domain that not only involves the GIG/DISN core network, but also the access and deployed tactical networks and interdomain traffic handling and SLAs. Different requirements for delay, packet loss and throughput validate the necessity of leveraging advanced IP-related mechanisms in different parts of the network, and using different architectures and mechanisms according to operator, while verifying the necessity for a standardized solution that respects the different implementations and still can satisfy the end-to-end needs.

QoS and SLA management will eventually become a mature technology that can be applied in the GIG/DISN networks. As QoS parameters and SLA/SLS definition becomes better understood and defined, SLA management and assurance will be implemented in more and more networks. SLA management seems to still be in the research area because high level automation of system correction is difficult to demonstrate: deployed and tactical network represent a case where the number of state variables is very huge, and consequently the control process is difficult to design. A first implementation of the SLA management may be first a tool providing a set of corrective solutions with the final decision left to the service provider.

References:

1. "DISA GIG Convergence Master Plan" Defense Information Systems Agency (DISA), 29 March 2006.
2. Center for DISN Services, "DISN Service Level Agreement for the Defense Information Systems Agency and its customers"
3. R. Braden, D. Clark, S. Shenker "Integrated Services in the Internet Architecture: an Overview", RFC1633 - IETF, June 1994.
4. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss. "An Architecture for Differentiated Service", RFC2475 - IETF, December 1998.
5. D. Grossman: New Terminology and Clarifications for Diffserv. RFC 3260. April 2002
6. F. Le Faucheur et al: Requirements for support of Diff-Serv-aware MPLS Traffic Engineering. RFC 3564. July 2003.
7. E. Rosen, A. Viswanathan, R. Callon "Multiprotocol Label Switching Architecture",RFC3031 - IETF, January 2001.
8. ETSI EN 301 790, "Digital Video Broadcasting (DVB); Interaction Channel for Satellite Distribution Systems", March 2003.
9. Prepared on behalf of the WiMAX Forum, "Mobile WiMAX – Part I: A Technical Overview and Performance Evaluation," February 2006.
10. D. Goderis and al., " Service Level Specification Semantics and Parameters ", -IETF Draft- <draft-tequila-sls-00.txt>, November 2000. http://www.ist-tequila.org