Government of Canada Gouvernement du Canada

700TAG – Technical Advisory Note #5

Congestion Management Techniques for Public Safety Mobile Broadband Communications Networks

700 MHz Mobile Broadband for Public Safety - Technology Advisory Group
Public Security Science and Technology

v.05b December 1, 2011

Partners: The Technology Advisory Group for 700 MHz Public Safety Spectrum (700TAG) is composed of a collaborative group of technical experts led by Centre for Security Science and includes scientific authorities from the Communications Research Center, and technical experts from Federal/Provincial/Territorial/-Municipal agencies.

Objectives

At times, wireless communication networks can become loaded to a point where performance degradation occurs, which can have a significant impact on the effective delivery of applications. This results when the available capacity is not sufficient to support traffic demands on the network. The Objective of this Technical Advisory Note (TAN) is to inform the Canadian public safety community on various techniques that can be used to mitigate these instances of network performance degradation. The action of each technique is described as well as the impact on the users' ability to make effective use of the network. Advantages and disadvantages of each technique are discussed. The TAN concludes with an assessment of how Congestion Management, while necessary to handle network performance degradation, constrains incident-response planning and execution.

Why is Congestion Management (CM) necessary?

The term “Congestion Management” in the context of communications networks refers to the application of processes or techniques that prevent a chaotic degradation of the perceived performance of the network when the demand for services exceeds the ability of the network to deliver those services. In fact, it’s not the network that degrades under the condition of traffic congestion, it is the applications, which are served by the network, that don’t respond as expected and thus, the user-experience can become unacceptable. Degraded performance can range from slow response to requests for information (lag) to loss of data which can manifest itself as distorted video and unintelligible audio for example

Figure 1 is a conceptual illustration of an end-to-end communications network – from the information consumer to the information source. The public safety users’ requests for information are carried across the local network and the high-capacity backbone network. The high-capacity backbone † would interface with other public safety networks and, possibly, commercial carrier networks. Firewalls and encryption protect the privacy of the data pertaining to each operator thus enabling multiple operators to carry their data on the same physical data pipe. The databases would generally be hosted on high performance servers residing in data centers that have high-capacity connections to the backbone network, which can be optical fiber-based.

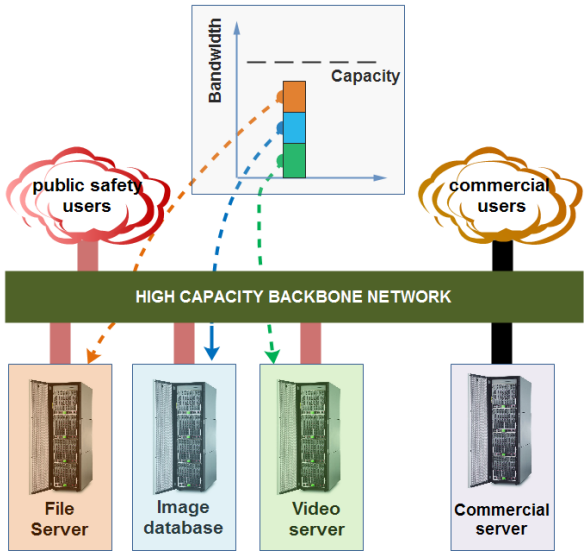

Figure 1: Conceptual illustration of an end-to-end communications network showing user requests for information carried across the High Capacity backbone and routed to the destination servers.

† An example of a high-capacity backbone network is the Government of Alberta’s “Supernet”. http://www.servicealberta.ca/1561.cfm

Figure 1 shows three users simultaneously sending requests for information. User "A" requests a data file, user "B" requests an image file, and user "C" requests a streamed video file. It is assumed that there is sufficient capacity in the local network and the backbone network to carry the three requests simultaneously.

Figure 2 illustrates an example of the response to the three user requests. The servers receive the requests and reply by transmitting the information. Typically, there would be more databases serving the public safety users than the three which are shown in Figures 1 and 2. This example assumes that the local network is capacity-constrained and that the throughput requirement of the three files, sent simultaneously, exceeds the instantaneous available capacity.

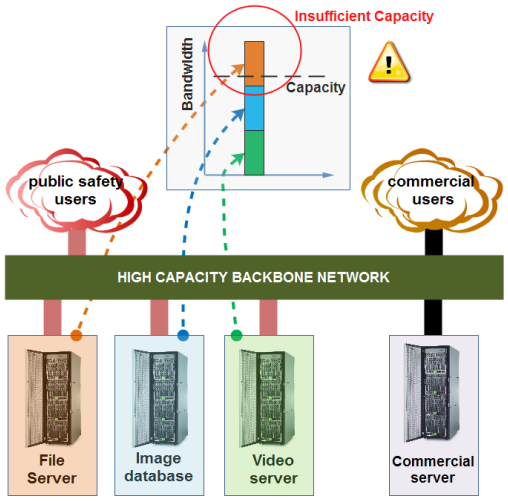

Figure 2: Illustration of servers responding to user requests for information.

In the example of Figure 2, there is insufficient capacity to pass all the packets at the same time. In a typical network, the edge routers will drop packets randomly if no congestion management exists. The video stream will be corrupted at the receiver and some of the other packets will be delayed due to buffering and re-transmission.

Mechanisms of Congestion Management

Fundamentally, CM means that human decisions (policies) need to be made on the relative importance of information and relative importance of users for particular circumstances. This in turn will be factored into the machine algorithms to determine what to do with packets when the channel is congested. This section of the TAN describes some of the enabling technologies that allow CM policies to take effect. Congestion Management techniques can be network-level CM or application-level CM. Both are used.

Link-by-link prioritization

The Internet Engineering Task Force (IETF) is responsible for developing standards that, among other benefits, ensure the continued interoperability of the Internet ecosystem. It has defined CM mechanisms at the network layer. The header of IP packets contain bits which can be used by Internet routers to prioritize some packets over others with 8 levels of granularity. Thus, ***Differentiated Services*** (DiffServ) [1,2] can differentiate traffic according to Type of Service (ToS). It can act on policies that would route high priority traffic (eg. real-time video) onto the least congested link before passing lower priority traffic (eg. emails). DiffServ can be used to treat latency sensitive traffic with higher priority. It is a rudimentary CM mechanism because its utility is limited to individual links between routers. End-to-end Quality-of-Service cannot be reliably assured on the basis of DiffServ.

Path prioritization

The IETF introduced ***Multi-Protocol Label Switching*** (MPLS) [2,3] for the Internet Protocol to emulate the switching efficiency of connection-oriented technology like Asynchronous Transfer Mode (ATM). An extension of MPLS, known as MPLS-TE allows end-to-end ‡ packet prioritization to be achieved with the potential for congestion avoidance. MPLS attaches a label to each packet, which is used to define the route that the packet will take through the MPLS-enabled backbone network from the ingress point to the egress point. MPLS also defines the Traffic Class (TC) of the packet. TC consists of 8 levels of priority for MPLS packets. The advantage of MPLS is that the MPLS-enabled routers can select the least congested paths for the packets, whereas DiffServ alone can only route packets along the least congested link to the next router. Figure 3 illustrates how MPLS-enabled routers can optimize the routing of packets in a way that DiffServ alone can not.

In the example of Figure 3, DiffServ will select the least congested link immediately following the ingress point, namely link AC. However, DiffServ is unable to factor the congestion in links CF and FG in its routing decision. MPLS, on the other hand, uses information about the degree of congestion along the entire path and, in this case,

‡ End-to-end means from the ingress to the egress points of the MPLS-enabled backbone network.

path AB-BE-EG would likely be selected. Routing decisions are made on a packet-by-packet basis since network congestion is highly dynamic and changing.

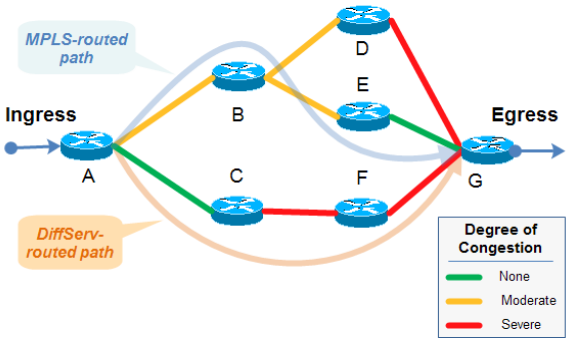

Figure 3: Illustration of QoS-aware end-to-end routing in an MPLS-enabled backbone network.

Last-mile Quality of Service

DiffServ and MPLS are routing protocols and have no effect on the Quality of Service (QoS) for packets in the access layer. That is, in the last-mile. Figure 4 illustrates the demarcation points between the high-capacity backbone network and the access layer. In this example, the access layer is the Radio Access Network (RAN).

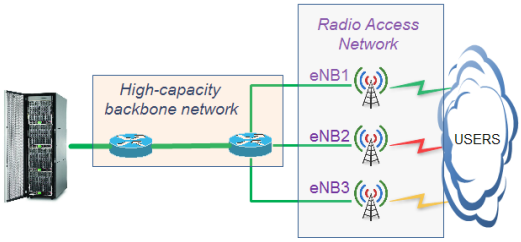

Figure 4: Demarcation points for the High-capacity backbone network and the Radio Access Network (RAN).

Note that in the example of Figure 4 the red highlighted link in the RAN (eNB #2) [§] is congested. MPLS congestion management via its label-switched routing and built-in QoS mechanism does not extend to the RAN. An additional level of CM is required in order to manage traffic in the RAN. In this section we will examine the CM capabilities of Long Term Evolution (LTE) since it is a candidate technology for the Public Safety Mobile Broadband Network (PSMBN), although CM techniques that can prioritize traffic and users could be implemented with other RAN technology by using external devices purposely designed for such use. These are commonly referred to as bandwidth managers or traffic shapers.

Access Control

LTE has built-in QoS and prioritization capabilities to manage congestion over the air-interface. The 3GPP [**] has specified fifteen ***Access Control Classes*** (ACCs) [4,5]. An example of how ACC can be applied to control access to the RAN is to bar non- public safety users, when circumstances warrant, from accessing the PSMBN. Of course, this assumes that commercial users are allowed onto the PSMBN.

LTE uses a prioritization scheme referred to as ***Allocation Retention Priority*** (ARP) [4]. ARP is a configurable attribute that is associated with the 15 ACC priority levels. Typically, the network administrator would define the ARP level for each user as part of the default setting in the user Profile database [††]. The ARP level is used to determine what action to take when there is insufficient capacity remaining in the eNB to satisfy a new user's request for air-link resources. There are 2 actions that can be taken: (i) deny the new user's request for access, or (ii) pre-empt an existing user in order to admit the new user.

Application QoS

A third CM scheme present in LTE is ***QoS Class Identifier*** (QCI) [4]. Whereas the two LTE schemes described above control the users' access, QCI assigns priority to applications according to 9 levels of QCI classifications. Three parameters define QCI: (i) priority, (ii) permissible latency, and (iii) maximum tolerable packet loss rate. The LTE resource scheduling functions allocate radio resources according to the priority of QCI classifications. The 3GPP has standardized the traffic-handling behaviour according to the assignment of QCI levels as shown in Table 1. The network operator can assign a QCI level to any application served by the PSMBN. The last column of Table 1 shows an example of how a network operator could associate QCI levels to applications.

§ eNB refers to the LTE Base Station.

** 3rd Generation Partnership Project (www.3gpp.org)

†† The User Profiles would be resident in a Subscriber Management database.

Table 1: Standardized QCI classifications for LTE [6].

QCI #	Priority	Delay Budget	Packet Loss Rate	Typical Application
1	2	100 ms	10^{-2}	Conversational voice
2	4	150 ms	10^{-3}	Conversational video (real-time)
3	5	300 ms	10^{-6}	Buffered streaming
4	3	50 ms	10^{-3}	Real-time gaming
5	1	100 ms	10^{-6}	IMS signaling
6	7	100 ms	10^{-3}	Live streaming
7	6	300 ms	10^{-6}	Buffered streaming, email, browsing, file download, file sharing, etc.
8	8	300 ms	10^{-6}	
9	9	300 ms	10^{-6}	

Asserting Congestion Management

In the previous section three technology enablers for CM were reviewed; (i) link-level prioritization, (ii) path prioritization at the network level, and (iii) access layer QoS. This section describes how CM policies are translated into action.

The process to convert CM policies into action is underpinned by the following aspects as illustrated in Figure 5:

- Policies and rules that dictate how a PSMBN should react when the demands exceed the available capacity.
- An understanding of how each network element in the PSMBN affects traffic in terms of how it shapes, throttles, routes, buffers, and prioritizes traffic.
- The purpose of the information that is carried over the PSMBN and its relationship to the applications that are invoked by users.
- The specific roles of the users when responding to incidents.

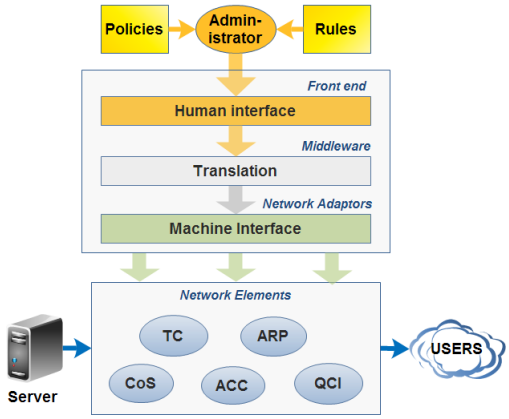

Figure 5: Converting Congestion Management policies and rules into action.

In reference to Figure 5, the Network Administrator of the PSBMN interacts with the PSMBN via a human-to-machine interface and enters congestion management instructions according to a prescribed manner (templates, forms, fields). The instructions represent the assignment of priorities to user profiles and to applications. The instructions could also specify conditions when different priority assignments should be used. For example, the conditions could be based on the severity, location, or type of incident. Another condition could be based on what agency has the lead role in responding to an incident.

Within the PSMBN is a middleware function that translates human-readable instructions into a set of command-level instructions (CLI). The CLIs are specific for each network element that affects the traffic flowing through the PSMBN. A human-level instruction would normally be translated into many CLIs since there would be several network elements to configure. Typically, a systems integrator would be responsible to develop the middleware.

In summary, CM action is accomplished when network elements are configured according to a set of instructions on what to do when the demand exceeds the capacity of the network. The emphasis is on (i) faithfully reproducing the instructions from policies and rules, and (ii) completely and accurately mapping the configuration points for each network element and understanding how they affect traffic as an integrated network.

Congestion Management Policies

Congestion Management is only required when demand exceeds capacity. At all other times CM does not affect the traffic flowing across the PSMBN. In LTE, CM policies apply on each bearer (service flow), not on the aggregate capacity. This section explores examples of the decisions that a network operator would need to make in order to manage data traffic. Relevant parameters for prioritizing traffic are also discussed.

The amount of traffic carried on the PSMBN is a function of three variables:

- Number of users,
- Data rate required by the applications,
- Correlation in time.

The capacity of the PSMBN is a function of:

- Amount of RF spectrum,
- Spectral efficiency of the wireless technology,
- Frequency reuse

The focus of this TAN is on the demand side and so the first approach to manage congestion would

be to minimize the demand wherever possible. The largest draw on bandwidth is video traffic. It requires high data rates, especially for high resolution, and the video sessions persist over long periods of time compared with the transfer of an image, an SMS message, or other types of data.

Taming video traffic

Video provides a rich source of information for situational awareness and so it is expected that a large number of sources of video will be present at an incident. The greater the amount of resolvable details required, the higher the video data rate needs to be. The amount of video traffic can be greatly reduced by limiting the number of feeds that are uploaded in high-resolution mode at any one time. This is made possible by switching the resolution at the source between high-res and low-res. High-res would be used only for those video feeds that are selected for viewing and low-res would be used for the rest. ***Decision-1***: *Limiting the number of video feeds is an additional consideration for incident planning. How many simultaneous high-resolution video feeds would be needed during a particular incident?*

Video traffic in the downlink direction can also be reduced by using multicast distribution. LTE supports enhanced-Multimedia Broadcast Multicast Services (eMBMS) which reduces the radio resources that are required to transmit the same video to multiple users. The penalty in terms of additional overhead is ≈10%, but a potentially larger penalty is that the eNB cannot use the Quality Channel Indicator feedback from the user equipment to optimize the performance in the presence of interference [7]. Modeling and simulation of specific deployment scenarios can assist with the trade-off analysis. ***Decision-2***: *Under what conditions would it be beneficial to use eMBMS?*

Prioritizing Users

The PSMBN is a finite resource shared amongst all the users. When there are too many users to satisfy all the demands one of the options available to the network operator is to limit the number of users. This can be done by assigning a priority level to each User in their user profiles. ***Decision-3***: *What set of criteria could the network operator use to establish who is assigned what level of priority?*

Ranking Applications

Some applications are more sensitive to latency and packet loss rate than others. Some are more important in terms of usefulness to the users for the mission at hand, while other applications may be more important for different missions. ***Decision-4***: *What set of criteria would the network operator use to establish which applications are more important? (eg. Which applications will be deemed to be mission-critical?)*

Parameters for prioritizing traffic

The premise for applying CM techniques is to introduce a degree of intelligence into the process of determining how to balance the services delivered to the users in those cases when all the demands cannot be met simultaneously. It is the "intelligence" programmed into the CM algorithms that ensures that critical applications are delivered to those users that have a critical need for it.

One of the bases for how users can be prioritized is according to the role of the user during the incident response. The Incident Command/Management System [8,9] lays out the roles that the responders would have in an incident and can be used to assign priorities to users. Another parameter can be the proximity or location of users relative to the incident since responders that are outside the incident area could be assigned lower priority in the eNBs that serve the incident. Yet another parameter could be the criticality of the applications for the incident at hand.

Depending on the incident some applications are more important than others. For example, accessing GIS and HazMat records may be more important at the scene of a fire than Automatic License Plate Readers or e-citation applications. But, during an Amber Alert Automated License Plate Readers would have higher priority. ***Decision-5***: *What factors must be correlated to impart "intelligence" into the PSMBN so that it will effectively manage congestion? How will the correlation rules be specified, implemented, and tested?*

Dynamic prioritization

The preceding sections introduce the notion that priorities are not static. Priorities are very much impacted by the incident. As such, a means is required to allow the Incident Commander or his/her delegate to tailor the priorities to an incident and to modify them as an incident unfolds in real-time. ***Decision-6***: *What decision-aiding tools will the Incident Commander need to be able to make the right prioritization decisions and in a timely manner in light of the flood of information that the PSMBN will offer him/her? How can response profiles that are pre-determined according to the type of incident be used? What kinds of levers will the Incident Commander need to adjust the priorities in real-time?*

Impacts of Congestion Management on Interoperability

The issues that CM attempts to resolve arise due to peaks in demand. Such peaks occur as a result of incidents where a larger-than-usual density of users converge and rely on situational awareness and other information in order to carry out their mission. Therefore, CM action is localized to an incident area, and if desired, could be managed locally by an Incident Commander. In this context, interoperability is primarily a matter of having a common method for applying prioritization criteria and conditions to all the responders, including those that are brought in from other jurisdictions to lend assistance.

In the case of mutual-aid where responders from other regional networks are under the same incident command as the local responders, interoperability requires that the users can be authenticated on the visited network. Once authenticated, the visiting users' priorities can be managed locally. However, since the Access Control Class of a user is stored in the USIM ‡‡ of the user equipment it is necessary that the regional network operators adhere to standard definitions of the Access Control Classes [4,5].

The technical capability is but one lane of the Interoperability Continuum [10]. It is equally important that the visiting responders be able to fit operationally into the incident organization. Therefore, if all regional networks use similar approaches to CM then the visiting users would have similar experiences concerning how the network responds to them, whether they are in their home network or visiting other networks.

Conclusion

There will always be incidents where there is insufficient capacity in the communications network to deliver the information that First Responders need to help them execute the mission at hand in the most effective manner. In some cases, the congested network will manifest itself by a slower than expected download of information. In other cases congestion may result in corrupted information especially for information that only has value for the users if in real-time. It is therefore, necessary to apply measures to mitigate the deleterious effects of congestion by controlling the response of the PSMBN in a predictable manner. This TAN examined the technical enablers of CM, how policies and rules for CM can be asserted by the network elements that shape traffic, and examples of the decisions that a network operator faces in order to establish an effective CM strategy.

As long as a user is authenticated on a PSMBN, his/her priority among the responders can be managed while operating in home networks as well as while visiting other regional networks. Interoperability is supported by the users' seamless experiences when visiting other networks, and reinforced by continual training and refining of the CM policies.

Limited capacity of a communications network imposes choices on First Responders. The less capacity there is in the network, the greater the degree of compromise that must be made. Congestion Management provides First Responders, incident planners, and Incident Commanders a way to make intelligent choices in the use of a finite resource – the RF spectrum.

An important objective of having a communications network whose behaviour is predicable during periods of unpredictable congestion is to gain the users' trust that they can count on this technology in critical circumstances. If some jurisdictions choose to implement the PSMBN with private partners, then that requires a high degree of transparency into the prioritization algorithms that the private partner implements on behalf of public safety and the ability to verify that what is promised is actually delivered.

‡‡ USIM is Universal Mobile Telecommunications Service Subscriber Identity Module.

References

Mode detailed information can be found in the documents referenced below.

1. Internet Engineering Task Force, "An Architecture for Differentiated Service", RFC2475, 1998.
http://datatracker.ietf.org/doc/rfc2475/?include_text=1

2. "Quality of Service Concepts", Cisco IP Solution Center Quality of Service User Guide, issue 3.0, OL-4345-01, Chapter 1
http://www.cisco.com/en/US/docs/net_mgmt/ip_solution_center/3.0/qos/user/guide/concepts.pdf

3. Yassine Hadjadj-Aoul, "Towards AQM Cooperation for Congestion Avoidance in DiffServ/MPLS Networks", Recent Patents on Computer Science, 2009.
http://www.benthamscience.com/cseng/samples/cseng2-1/0001CSENG.pdf

4. Wim Brouwer, "QoS in LTE – PSCR Demo Days", Alcatel-Lucent, Dec.2010
http://www.pscr.gov/projects/broadband/700mhz_demo_net/stakeholder_mtg_122010/day_1/5.2_qos_priority_preemption-alu.pdf

5. 3GPP TS22-011, "Technical Specifications Group Services and Systems Aspects – Service Accessibility (Release 8)", Chapter 4 – Access Control, V8.9.0, Sept.2009
http://www.3gpp.org/ftp/Specs/html-info/22011.htm

6. Harri Holma and Antti Toskala, "LTE for UMTS – OFDMA and SC-FDMA Based Radio Access" John Wiley and Sons, 2009.

7. Ozgur Oyman and Jeffrey Foerster, Intel Corporation, Yong-joo Tcha and Seong-Choon Lee, KT Corporation, "Toward Enhanced Mobile Video Services over WiMAX and LTE", IEEE Communications Magazine, August 2010, pp.68-76.

8. US Department of Homeland Security, "National Incident Management System – Appendix B: Incident Command System", December 2008.
http://www.fema.gov/pdf/emergency/nims/NIMS_AppendixB.pdf

9. Government of Ontario - Ministry of Community Safety and Correctional Services "Incident Management System for Ontario", December 2008.
http://www.emergencymanagementontario.ca/english/ims/ims_main.html

10. "Communications Interoperability Strategy for Canada", Public Safety Canada.
http://www.publicsafety.gc.ca/prg/em/cisc-eng.aspx

NOTE: *DRDC Centre for Security Science warrants that this advisory note was prepared in a professional manner conforming to generally accepted practices for scientific research and analysis. This advisory note provides technical advice and therefore is not a statement of endorsement of Defence Research Development Canada, Department of National Defence, or the Government of Canada*

Author: Claudio Lucente, P.Eng.

Scientific Authorities:
Jack Pagotto, Head/ESEC S&T [Emergency Mgmt Systems & Interoperability, Surveillance/Intel, E-security, Critical Infrastructure Protection]
Joe Fournier, CRC

Approval for Release: Dr. A. Vallerand