Issues on the Allocation of Performance Objectives for Common Channel Signaling (CCS) Networks

Frank Y.S. Lin
Bellcore
Room RRC-4A1009
444 Hoes Lane
Piscataway, NJ 08854-4182
(908) 699-4220 / fax (908) 336-2275
lin1@cc.bellcore.com

Marco M. Mostrel
Bellcore
Room RRC-4A1009
444 Hoes Lane
Piscataway, NJ 08854-4182
(908) 699-8640 / fax (908) 336-2275
marco@cc.bellcore.com

Abstract

In this paper, we discuss issues on the allocation of end-to-end performance objectives to Common Channel Signaling (CCS) network elements. Practical considerations result in end-to-end performance objectives being usually allocated to network elements in such a way that, if the allocated performance objective for each network element is satisfied, then the end-to-end performance objective can be guaranteed. This allocation process makes it possible to plan and engineer individual network elements according to uniform rules. It is not trivial to determine a feasible allocation scheme, let alone an effective one, especially when percentile types of end-to-end performance objectives are involved, as is the case for CCS networks. This paper outlines a number of key issues, and proposes possible solutions to some of them.

1. INTRODUCTION

Allocation of performance objectives consists of determining individual network element performance objectives, such as cross-network element delay or linkset utilization, given a desired global performance objective, such as end-to-end delay for a Common Channel Signaling (CCS) network signaling section. Characteristics of the CCS traffic and a variety of CCS network element failure scenarios under which these objectives should be achieved need to be considered. A notable example of a global performance objective is the regulatory performance objective of end-to-end call setup delays under five seconds for the 800 database service that is mandated by FCC Docket No. 86-10 [1].

Traditionally, for engineering tractability, a feasible approach is to allocate end-to-end performance objectives among individual network elements in such a way that, if the individual performance objective for each network element is satisfied, then the end-to-end performance objective can be guaranteed. Individual CCS network elements (and the office equipment and interoffice facilities they ride on) can thus be engineered according to uniform rules. Performance objectives can also be used by CCS network planners for sizing and capacity expansion of CCS network elements.

For each service relying on CCS networks for signaling, performance objectives need to be established. Allocation of multiple, concurrent, performance objectives is therefore involved. In such a multi-objective environment, different types of CCS traffic (and, possibly, different priority queueing disciplines) need to be considered. For example, in addition to the FCC 86-10 800 call setup delay regulatory objective mentioned earlier, one needs to consider delay objectives for other services, such as Advanced Intelligent Network (AIN) services. This imposes additional constraints on the allocation schemes, and might require that allocation schemes be reviewed when a new service is introduced.

Two categories of performance objectives are important for the CCS network: availability or downtime objectives, and delay or utilization objectives. These two categories are discussed in Sections 2 and 3, respectively. Section 4 discusses queueing models for CCS network elements, and how they can be used to help model performance objectives and the allocation thereof.

2. AVAILABILITY AND DOWNTIME OBJECTIVES

Availability or **downtime** objectives are intended to control the amount of time the CCS network (or portion thereof) is able to perform its required function. They are typically measured by a single number equal to the long-term percentage of time the CCS network – or portions thereof – are expected to be "down". The expected percentage of downtime for a network element can be interpreted:

— either as the average downtime over many years for this network element,

— or as the average downtime over one year for a population of the network elements.

As such, availability objectives can significantly influence end user perception of service quality.

An end-to-end downtime objective for each CCS network signaling section was established the American National Standard ANSI T1.111.6-1992 [2] at ten minutes per year. Based on this objective, downtime allocations to the access portions and the backbone portions were set respectively to

0-7803-1820-X/94 $4.00 © 1994 IEEE

five minutes per year (three minutes per year of service switching point (SSP) downtime + two minutes per year of A-linkset/home signaling transfer point (STP) downtime), and zero minute per year (i.e., negligible downtime for B-/D-linkset quads and C-linksets). Downtime of the access portion of a CCS network signaling section causes isolation of the end office served; it occurs when a pair of A-linksets is severed. Downtime of the backbone (quad linksets and C-linkset) portion of a CCS network signaling section causes isolation of the four STPs and all the end offices served; it can occur only with multiple failures.

These downtime performance objectives are very stringent and might not always be met. Improving the reliability of the network components used, increasing the dependability of the software (for example by using software diversity for the generics used in a mated pair of STPs), enhancing the diversity of the CCS network architecture selected, and provisioning interoffice routes to support the access and backbone linksets that use physically diverse facilities and equipment sites, can help CCS network planners design a network that will perform closer to these downtime performance objectives. The reader may consult References [3] for further reading on various route diversity strategies which can be used to increase CCS network survivability, and [4] and on a four-layer framework for describing and comparing survivability techniques.

3. DELAY AND UTILIZATION OBJECTIVES

Delay objectives have the most direct impact on CCS traffic engineering. Three major types of delay objectives (as well as hybrids of these types) are usually considered: peak (or maximum), percentile, and mean (or average). A peak (or maximum) delay objective requires that messages be transmitted within a given time period. A percentile delay objective requires that a certain percentage of all messages be transmitted within a given time period. A mean (or average) delay objective requires that, on the average, messages be transmitted within a given time period.

Network element **utilization** is related to cross-network element delay. For a signaling link, it is defined as the fraction of time packet transmission occurs over the link. It can be computed as the ratio of the link carried load to the link capacity (or link speed). Processor utilization can be computed in a similar fashion. Low utilization implies higher costs. High utilizations, on the other hand, may cause unacceptable delays. Even if the network element is engineered for a low utilization, CCS network element failures might cause some additional signaling traffic to be diverted to the network element, thereby increasing (usually, doubling for a single network element failure) its carried load and its utilization. Utilization upper bounds must therefore be found to provide both congestion control and survivability at reasonable costs.

3.1 Time Scales Involved in Delay or Utilization Objectives

Performance objectives such as delay or utilization objectives can be specified as peak (maximum) over a given time interval, or mean (average) over a given time interval, or else are relative to the completion of a given type of call. Whether peak or mean delay/utilization objectives are appropriate, and how long the time intervals over which these statistics should apply must be determined.

For example, a requirement in the Bellcore technical specification of SS7 [5] aims at a signaling link load of 0.4 erlangs (i.e., utilization objective of at most 40% as originally recommended in Reference [6] and later adopted in Reference [5]), so that if a failure occurs, an expected peak load of at most 0.8 erlangs[1] of CCS traffic would be carried by the surviving signaling link. The time scale involved in this utilization objective is however left undefined in Reference [5]. It is not clear whether 5-minute or hourly averages are needed to adequately capture the variations in carried loads.

For the Public Switched Telephone Network (PSTN) for example, traffic engineering is based on performance objectives which involve maximum busy-season, busy-hour blocking. More study of the characteristics of CCS traffic – which is likely to be burstier than PSTN traffic – is needed, in order to determine the relevant time scales for the CCS network. The reader is referred to Reference [7] for ongoing work on this issue.

3.2 Allocation Schemes for End-to-end Delay Objectives

Even when end-to-end delay objectives are defined, allocating them to individual cross-network element delays is difficult. For example, it is not clear how the 5-second call setup delay for the 800 database service should be apportioned to the call segments: Plain Old Telephone Service (POTS) connection to switch, intra-SSP processing, access to backbone CCS network, interconnection to interexchange carrier (IC), access to service control point (SCP), etc. When a hybrid delay objective involving the peak, percentile and mean criteria is considered as in the FCC 86-10 case[2], finding an effective allocation policy is a difficult task. Criteria/rules are needed to effectively perform this allocation of delay among CCS network elements. For example, one may want to increase the utilization of (and thus allocate higher admissible delays to) more expensive CCS network elements.

---

1. This is based on the assumption that unacceptable delays occur when the load on a signaling link goes over 0.88 erlangs. At 56 kbps, this corresponds to a maximum link capacity of 0.88*56,000/8 = 6,160 octets per second. The 0.8 erlangs peak load is obtained by subtracting a security margin of about 10% from the 0.88 erlangs value.
2. The FCC 86-10 ruling requires that (i) 97% of the 800 traffic involve call set-up times of five seconds or less as of May 1, 1993, and (ii) 100% of the 800 traffic involve maximum call set-up times of five seconds or less and average call set-up times of 2.5 seconds or less by March 4, 1995.

Allocating a percentile type of delay objective is more difficult than allocating an average (mean) type of delay objective. In a recent study on Switched Multi-megabit Data Service (SMDS) networks [8], percentile end-to-end delay objectives were considered, and a number of approaches for allocating these delay objectives have been proposed and evaluated. Among the proposed approaches, the *G1/G/1* bound approach [9] seemed particularly promising due to its generality, low complexity, and the good bound quality shown in the computational experiments. Given the packet interarrival time and service time distribution (more precisely, the Laplace transform) on a network element, the $G1/G/1$ bound approach can be used to calculate an exponential bound (from above) on the tail of the equilibrium waiting time distribution. Based upon a theoretical result and a closed-form inverse Laplace transform presented in Reference [8], one can calculate an upper bound on the maximum link utilization in a path with multiple links. One may investigate the feasibility of applying these results to CCS networks.

3.3 Utilization Objectives to Protect Against Specific Failure Events

The utilization objective guideline of 40% applies to all linksets; namely, A-, B-, C-, D-, E- and F-linksets should be engineered so that their utilization is at most 40% under normal (no failure) conditions. Moreover, C-linksets carry CCS traffic only in case of B-/D- quad or STP failure (they also carry synchronizing – and, for some vendors, proprietary – signals between STP mates).

The 40% utilization objective aims at providing survivability in the event of a single failure of the access network, or a single failure of the backbone network, or most double failures of the backbone network. In the event of such a failure, congestion is avoided if the utilization remains below the critical value of 80%. However, the backbone portion of a CCS network engineered according to this requirement cannot withstand a double failure such as the simultaneous failure of an STP and one of the two B- or D- quad linksets connecting its mate to the other mated pair of STPs [3] To remedy vulnerability to this type of double failure, it has been suggested to reduce the utilization objective for B-/D- quad links to a value below 40%; for example, 20%.

Another requirement in the Bellcore technical specification of SS7 [5] is that an STP should be able to handle its mate's traffic load in addition to its normal (no failure) traffic load. No similar requirement for CCS nodal network elements other than STPs is given in Reference [5]. SCPs and databases might be candidates for similar mating requirements.

4. QUEUEING MODELS FOR CCS NETWORK ELEMENTS/SYSTEMS

Any CCS network element (SSP, STP, linkset) or CCS network system (SCP) can be viewed as a queueing system, where the processor (link transmitter in the case of a linkset) plays the role of server, and the message packets play the role of customers to be served. Queueing theory can therefore be used to model cross-network element/system delays, and end-to-end delays. More theoretical study should be devoted to determining the appropriate queueing models needed to approximate the delay vs. utilization curve for CCS network elements/systems, so that utilization bounds such as the 40% guideline in Reference [5] can be refined, and expected end-to-end delays can be better estimated for a variety of CCS network signaling sections.

One major difficulty associated with this approach is that the interarrival time and service time distributions are usually not available from traffic measurements. In addition, even though the interarrival time and service time distributions are known, it may be difficult to exactly calculate the average packet delay. As a result, approximation/estimation of the distributions and the average delay may be needed. Sensitivity of delay with respect to approximate/estimated distributions must be investigated. What follows is a more detailed discussion of a number of issues when queueing models are applied.

4.1 Characterization of the Packet Arrival Process

Can the interarrival time be adequately characterized by a certain distribution? What parameters are needed? A recent study [10] shows that Weibull and hyperexponential distributions are suitable candidate for characterizing Local Area Network (LAN) traffic. To determine whether the interarrival times fit a hypothesized distribution, one can perform standard goodness-of-fit tests using, for example, the Kolmogorov-Smirnov statistic [11] or the Anderson-Darling statistic [12] (for normal or exponential distributions), to obtain a quantitative measure of how close a set of samples agrees with the hypothesized distribution. The issue then becomes what level of significance (e.g., 5%, 1%?) should be used for engineering purposes.

Recent studies suggest that the packet arrival process to an originating STP might be close to (no measure of how close is given) a Poisson process [13]. If most of the CCS traffic consists of call setup packets, then the Poisson assumption seems to be valid, but only for short engineering periods, e.g., five minutes, since the average demand changes with time. The stationary property, however, probably does not hold if longer time periods are considered, e.g., one hour, because then the coefficient of variation of the interarrival time is greater than 1. To illustrate this situation, consider the following example of a one-hour period of CCS traffic where the packet arrival process for each of the twelve 5-minute intervals T_i ($i = 1,2,...,12$) is assumed to be Poisson with mean μ_i. Assume that $\mu_i \neq \mu_j$ for a pair (i,j). An analysis of the samples of interarrival time in the one hour interval (without considering dependency among samples) may indicate that the overall interarrival time distribution can be characterized by a hyperexponential distribution with 2 or more stages. Consequently, the coefficient of variation is greater than 1. If an $M/G/1$ queueing model with the average arrival rate equal to $\sum_{i=1}^{12} \mu_i/12$ for the one-hour interval is used, the average

packet delay will be underestimated. This can be shown as follows. Consider the mean arrival rate μ as a random variable. It can be shown that the average packet delay from the Pollaczek-Khinchin (P-K) formula [14] is a convex function of μ. Then by Jensen's inequality the result follows.

4.2 Characterization of the Packet Service Time Distribution

Besides the traffic arrival process, one also needs to study the distribution of packet service times. Can the packet distribution be evaluated from traffic measurements? If not, given the possible range (upper and lower bounds), mean of packet service times, and mean interarrival time distribution, what is the packet service time distribution that leads to the maximum delay? This is a worst-case analysis with imperfect information and, therefore, conservative objective allocation policies will be obtained. Another question to be answered is whether the service time distribution varies with time.

4.3 Characterization of Unobservable Queues

When a network element/system's input process is not observable, can one infer it from the characteristics of upstream queues? This issue is important since some network elements/systems have complex internal structures and can be modeled as a network of queues. However, in some cases, one can only observe the traffic flows into and out of the network element/system, but cannot observe a critical/dominant internal queue, to investigate the traffic characteristics.

4.4 Kleinrock's Independence Assumption

Kleinrock's well-known independence assumption [15], which states that merging several packet streams on a transmission line has an effect akin to restoring the independence of interarrival times and packet lengths, may not be valid for a small number of merging traffic streams, as for example in a standard B- or D-linkset quad configuration, especially under certain failure scenarios (e.g., an STP fails). Simulation has shown [14] that under heavy traffic conditions, the average packet delay is smaller in the (real) case where packet interarrival times and service times are correlated than in the case where the independence assumption holds. The reverse holds true when traffic is light. It is not known whether and in what form this result can be extended to more general networks. Simulation work is needed to evaluate the validity of the independence assumption for CCS traffic.

4.5 CCS Traffic Characteristics

Characteristics of the CCS traffic need to be properly understood; in particular, one needs to determine what statistics (mean, variance and/or higher moments) of the packet interarrival time distribution and the service time distribution are essential to properly model CCS traffic. Appropriate traffic sampling methods and the most relevant traffic engineering periods must be determined.

In addition to mean delays, **burstiness** might also need to be considered. for an analysis of burstiness in computer network traffic). Burstiness describes the variability of the interarrival time distribution. It can be defined in various ways, e.g., variance, peak-to-mean ratio, variance-to-mean ratio (peakedness), or coefficient of variation. Therefore, one parameter, i.e., mean, may not be sufficient to characterize traffic, as it is for exponential distributions. An appropriate definition of burstiness must be determined so that the characteristics of traffic are best described, while at the same time, the parameters are easy to measure. For example, if we define burstiness as the coefficient of variation of the interarrival time, then calculation of the first two moments of the interarrival time is required. The second moment, however, need not be measured on a regular basis, so that only the maximum burstiness is estimated.

Two other issues may arise with the definition of burstiness. Assume burstiness is defined as the coefficient of variation of the interarrival time. Two different interarrival time distributions, such as the distribution involved in "bulk arrival" systems and the hyperexponential distribution, may have the same burstiness (since they have the same first two moments). Their burstiness is greater than 1. However, these two distributions may lead to very different delays. It would be useful, therefore, to determine the largest (worst-case bound) mean delay realized among interarrival time distributions (where the service time distribution is fixed) with given mean and burstiness. Additionally, the value of burstiness is dependent on the time scale chosen. If a Poisson arrival process without the stationary property is considered, burstiness is closer to 1 for short engineering periods than for longer ones (see the example in Section 4.1).

4.6 Kingman's Approaches to Upper Bound Delays

If it turns out that the arrival process is not Poisson and hence the $M/G/1$ queueing model is not valid, two approaches proposed by Kingman can be used to calculate upper bounds on average delay for $G/G/1$ queues. Using Kingman's first approach [16], an upper bound on the average waiting time of a $G/G/1$ queue can be calculated, given the first two moments of the interarrival time and the service time. This approach has the following advantage. First, the bound is easy to calculate. Second, one can calculate a feasible region of the mean and variance of the interarrival times so that the delay objective can be guaranteed. This can provide an easy way to engineer the traffic and/or system capacity. For example, if one knows the largest coefficient of variation of the interarrival times and the acceptable average delay on a network element/system, then a threshold on the system utilization can be calculated. Third, the bound improves as the utilization factor increases. The high utilization region is where traffic and capacity engineering are most important and sensitive. The applicability of the above approach requires further investigation.

If the packet interarrival time distribution and service time distribution are known, sharper upper bounds on the average delay may be obtained using another result by Kingman [9], whereby the worst-case (given the mean and burstiness/range) distributions mentioned earlier to overestimate the average

delay are used. The relative effectiveness of Kingman's two approaches needs to be investigated.

5. SUMMARY

Two categories of performance objectives are important for the CCS network: availability or downtime objectives, and delay or utilization objectives. Performance objectives are needed by CCS network planners and engineers for monitoring, sizing and expanding capacity of CCS network elements/systems.

Given a desired global performance objective, such as end-to-end delay for a CCS network signaling section, allocation of this performance objective consists of determining individual network element/system performance objectives, such as cross-network element/system delay or linkset utilization, in such a way that, if the individual performance objective for each network element/system is satisfied, then the end-to-end performance objective can be guaranteed. Individual CCS network elements/systems can then be planned and engineered according to uniform rules.

In this paper, we have discussed issues including: (1) the difficulty to achieve end-to-end downtime objectives, (2) the need to clarify time scales involved in delay or utilization objectives, (3) the need for efficient allocation schemes for end-to-end delay objectives, (4) the need to determine utilization objectives to protect against some double failure events, (5) the need to investigate queueing models for CCS network elements, and (6) need to determine performance objectives for new services. For a number of issues discussed, possible solutions have been proposed.

The emphasis of this paper was (1) to address the importance of properly allocating end-to-end performance objectives to CCS network elements/systems, and (2) to highlight issues involved in this allocation process. We also identified areas for further investigation. Satisfactory resolution of these outstanding issues may lead to more efficient and effective allocation schemes, and help improve the network planning and traffic engineering process for CCS networks.

REFERENCES

1. *"Provision of Access for 800 Service"*, Memorandum Opinion and Order on Reconsideration and Second Supplemental Notice of Proposed Rulemaking, CC Docket No. 86-10, Federal Communications Commission, Washington, D.C., 8/1/91.
2. American National Standard ANSI T1.111, *"Signalling System Number 7 (SS7) – Message Transfer Part (MTP)"*, Section 5.1.2 (1992).
3. Mostrel, M., *"Issues on the Design of Survivable Common Channel Signaling Networks"*, IEEE Journal on Selected Areas in Communications, Vol. 12, No. 3 (1994), pp. 526-532.
4. *"A Technical Report on Network Survivability Performance"*, T1 Technical Report No. 24, Committee T1 – Telecommunications (1993).
5. *"Bell Communications Research Specification of Signaling System Number 7"*, Bellcore Technical Reference, TR-NWT-000246, Issue 2, June 1991, Revision 2, December 1992.
6. Ahmadi, H., and Akinpelu, J., *"Recommendations for CCS Link Engineering"*, AT&T Bell Laboratories Internal Document, December 1983.
7. Duffy, D., McIntosh, A., Rosenstein, M., and Willinger, W., *"Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks"*, IEEE Journal on Selected Areas in Communications., Vol. 12, No. 3 (1994), pp. 544-551.
8. Lin, F.Y.S, *"Allocation of End-to-end Delay Objectives for Networks Supporting SMDS"*, Proceedings of IEEE Global Telecommunications Conference, pp. 1346-1350, November 1993.
9. Kingman, J., *"Inequalities in the Theory of Queues"*, Journal of the Royal Statistical Society, Series B, Vol. 32, pp. 102-110, 1970.
10. Flaki, S.O., and Sorensen, S.A., *"Traffic Measurements on a Local Area Computer Network"*, Computer Communications, Vol. 15, No. 3, pp. 192-197, April 1992.
11. Trivedi, K.S., *"Probability & Statistics with Reliability, Queueing, and Computer Science Applications"*, Prentice-Hall, 1982.
12. Davis, C.S., and Stephens, M.A., *"Empirical Distribution Function Goodness-of-fit Tests"*, Applied Statistics, Vol. 38, No. 3, pp. 535-582, 1989.
13. Kagan, J.S., and Weingarten, A., *"Analysis of CCSN/SS7 Link Traffic Engineering Algorithms"*, Proceedings of IEEE Global Telecommunications Conference, pp. 1735-1740, December 1992.
14. Bertsekas, D., and Gallager, R., *"Data Networks"*, Prentice-Hall, Inc., 1987.
15. Kleinrock, L., *"Communication Nets: Stochastic Message Flow and Delay"*, McGraw-Hill, New York, 1964.
16. Kingman, J.F.C., *"Some Inequalities for the Queue GI/G/1"*, Biometrika, Vol. 49, pp. 315-324, 1962.