Use of Approximation Methods in the Analysis of Communication Networks with Heterogeneous Traffic

Dr. Sayeed Ghani
Chairman, Dept. of MIS & Computer Science
Institute of Business Administration, Karachi
sghani@iba.edu.pk

Abstract

The Convergence of various classes of traffic in modern Telecommunication networks has significantly increased the complexity of performance analysis of such networks. Exact solutions being computationally intractable, it is of importance to have analytical approximations that can be used in various key aspects of such networks such as control of traffic and optimization of resources. This paper looks at various approximation techniques used in the past and in particular a Decomposition Approximation that is valid when heterogeneous traffic are utilizing shared resources or channels, where the holding times of the traffic vary greatly. This paper briefly describes the basic approach of this method and gives examples of the use of such approximation techniques. Future applications of this method are also given where this method could be usefully applied.

1. Introduction

With the advent of heterogeneous networks carrying a mix of voice, data and video traffic, the complexity of performance analysis of communication networks has considerably increased. Most exact solutions of networks with integrated voice/data traffic are computationally intractable for large systems. Due to this increase in complexity, finding approximation methods which can be conveniently utilized for analyzing the key performance criteria of such networks has become increasingly important.

Heterogeneous networks are characterized by various classes of traffic that need to be transmitted over shared resources such as land-line, wireless or satellite channels. In most such access multiplexers the objective is the optimal utilization of resources such as bandwidth and buffer space, while minimizing cost functions based on Quality of Service (QoS) parameters such as blocking probability and delay. The access network frame structure typically is a combination of circuit switching and packet switching as in the case of GPRS/GSM networks (see [1], [2], [3] and [4]) or may consist of an ATM cell based structure (see [5], [6], [7] and [8]). Each of the classes of traffic generally have their own characteristics such as constant bit rate (CBR) or variable bit rate (VBR), and bandwidth, delay and jitter requirements.

In section II of this paper we consider a GSM/GPRS network with two classes of traffic: voice and data. Here, a brief overview of the approximation techniques which have frequently been applied to such networks is described, and the decomposition approximation is briefly reviewed and the conditions under which the approximation is valid are presented. Section III gives the GSM/GPRS system model and shows how the Decomposition Approximation can be applied to such a network. In section IV an alternate model of isochronous and asynchronous traffic using self-similar processes is considered and applications of the technique are reviewed here. In section V a traffic model in the case of ATM based networks is also reviewed where the approximation has been applied. Finally in section V conclusions are submitted and further areas in which the technique can be applied and suggestions for future research work are presented.

II. Voice/Data Integration in GSM/GPRS

The Global System for Mobile (GSM) along with General Packet Radio Service (GPRS) is an example of a network that has two classes of traffic, voice and data, competing for the bandwidth resources of the access network to the Base Station. GSM which provides a circuit switched infrastructure for carrying voice traffic is not an efficient mode of transport for carrying data traffic. Hence the GPRS standard from the European Telecommunications Standards Institute (ETSI) is designed as an extension to the GSM network, and allows for unused GSM channels to be utilized for GPRS data traffic, which are transmitted as packet switched data.

The traffic model generally assumes that a certain number of channels are dedicated for GPRS and the remaining channels are shared with GSM. In such a model we assume that N physical transmission channels are available in the system. Of these N_d are dedicated for GPRS data, and $N_v = N - N_d$ are shared between circuit switched GSM voice and GPRS data, with GSM voice having either preemptive or non-preemptive priority over GPRS data. This frame structure is depicted in Figure 1.

The GPRS data experience queuing in the access multiplexer and the size of the buffer is limited to a maximum buffer size. The GSM voice calls do not experience any queuing in the case of preemptive-priority, however in the case of non-preemptive priority over GPRS data a limited queuing may be experienced [9].

The above model is reminiscent of the early architectures for the multiplexing of voice and data in integrated voice/data networks such as the Senet (Slotted Envelope Network) based on a moveable boundary scheme [10], [11]. Such a model has been heavily studied with both exact methods [12], [13], and approximation methods with a number of variations including modeling voice with Digital Speech Interpolation (DSI) [14], [15].

The arrival process of voice and data are approximated by a Poisson arrival process, with average arrival rates of λ_v, λ_d respectively, and the holding times, neglecting the granularity of the frame structure, are approximated by an exponential distribution, with average holding time of μ_v, μ_d respectively. This results in a two-dimensional Markov chain depicting the voice/data process whose equilibrium stationary distribution is desired.

Exact techniques including matrix methods and moment generating functions require finding the roots of a polynomial within the unit circle. For large problems this poses convergence problems since the roots are generally close to one. Matrix techniques give rise to very large matrices since the state space can be very large, and hence can become computationally very expensive.

As a result various approximation techniques have been applied to this problem. Prominent among these are the fluid-flow [14], [15], [16], and diffusion [17] approximations, which have been applied to infinite data buffer systems. Hence these approximations are inherently limited to the heavy traffic regions. The fluid-flow approximation makes the assumption that the data flow is deterministic, i.e. data arrive at a constant rate and are also serviced at a constant rate. The justification of this approximation is that when voice holding times are much longer than data bursts, most of the queuing of data messages is due to fluctuations in the voice rather than fluctuations in the data traffic. With this approximation, the data process given the voice process is a deterministic process. The main computations in this analytic approach are the evaluation of the eigenvalues and eigenvectors of a matrix, and the solution of a set of linear equations [14]. Here, closed-form solutions can only be obtained for very small systems.

The Decomposition Approximation proposed in [18] and [19] essentially approximates the steady state behavior of such systems by 'decomposing' them into long and short term behavior. This results in the conversion of a multidimensional Markov chain into a hierarchy of groups of states, such that the interaction *between* the groups is mall compared to the interaction *within* the groups. Hence the short term equilibrium distribution of a group is approximated by ignoring its interaction with other groups.

This approximation is based on the assumption that each group of states should achieve equilibrium in isolation. In certain scenarios this is feasible such as for those in which the holding times of one class of traffic are much longer to that of another. This is typically true for voice/data integration, where voice holding times may range in the hundreds of seconds, and data transmission times may only require milliseconds, resulting in a ratio as large as 10^4. However, for infinite buffer systems, the above approximation becomes invalid if the system is in the overload region, i.e. where the average data utilization is high enough to require some of the shared channels for transmission [18].

A major advantage of this approximation technique is that it generally results in closed-form solutions. Such solutions do not require any computational complexity and are also very useful for control and optimization purposes (see e.g. [20], [5], [6], [7], [8]).

III. GSM/GPRS System Model

The GSM/GPRS system described above can be modeled by a Markov chain. In the case of GSM voice having preemptive priority over GPRS Data, the model is described by a two dimensional process: $\{V(t), D(t)\}$, here:

$V(t) = i$ = Total number of circuit switched GSM voice calls, $i = 0, \cdots, N_v$;

$D(t) = j$ = Total number of GPRS data packets in the system, $j = 0, \cdots, M$;

Here M is the maximum number of data packets that are allowed in the system. Due to the preemptive priority assumption, the voice process is independent of the data process, and it can hence be simply modeled as a $M/M/N_v/N_v$ queue. The probability of having i voice calls in the system, defined as $p_v(i), i = 0,1,\ldots,N_v$ is given by:

$$p_v(i) = \frac{\rho_v^{\,i}}{i!} \Big/ \sum_{j=0}^{N_v} \rho_v^j / j! \qquad \text{for } i = 0,1,\ldots,N_v$$

Here we have defined $\rho_v \equiv \lambda_v / \mu_v$ as the average GSM voice utilization. The performance criterion of interest for the GSM voice is the call blocking probability i.e. the probability that an arriving GSM voice call is not accepted into the system. This is then given by:

$$P_{BV} = p_v(N_v) = \frac{\rho_v^{N_v}}{N_v!} \Big/ \sum_{i=0}^{N_v} \rho_v^i / i!$$

However, the GPRS data process *does* depend on the voice process. The state transition diagram of the associated two dimensional Markov $\{V(t), D(t)\}$ is shown in Figure 2. A closed-form solution is, as was discussed earlier, not readily available. The Decomposition Approximation is now based on the following underlying assumption of [18] that α, defined as the ratio of GSM voice holding times to that of GPRS data, is large, i.e.:

$$\alpha \equiv \frac{1/\mu_v}{1/\mu_d} \quad \rangle\rangle \quad 1 \cdot$$

This is based on the fact that typical GSM voice calls occupy channels for 120 – 180 seconds; as compared to GPRS data packets which are typically transmitted within 2 – 10 seconds (see [1]).

The decomposition technique essentially assumes that the steady state behavior of such systems can be approximated by converting the multi-dimensional Markov chain into a hierarchy of group of *aggregate states*, such that the interaction *between* the groups is small compared to the interaction *within* the groups. A group of states would thus comprise all the states for a fixed number of GSM voice calls. Hence for the duration of a GSM voice call, the technique assumes that the GPRS data process achieves steady state, and hence its equilibrium distribution may be approximated by ignoring the transitions between groups.

The result is a series of decomposed data processes that can in isolation be modeled as a $M/M/N-i/M$ queue, where i the number of voice calls in the system can be assumed to be constant. The joint probability $p(i,j) = \Pr\{V(t) = i, D(t) = j\}, i = 0,1,\ldots,N_v, j = 0,1,\cdots,M$ can then be approximated by:

$$p(i,j) \approx p(j \mid i) \times p_v(i), \qquad i = 0,1,\ldots,N_v, j = 0,1,\cdots,M$$

Here, $p(j \mid i) \equiv \Pr\{D(t) = j \mid V(t) = i\}$ is the conditional probability of GPRS data being in state j, given that GSM voice is in state i. The conditional probabilities are then given as follows [18]:

$$p(j \mid i) = \begin{cases} P_{i0}\rho_d^j / j! & 0 \le j < N-i \\ P_{i0}\dfrac{\rho_d^j}{(N-i)!(N-i)^{j-(N-i)}} & N-i \le j \le M \end{cases}$$

$$P_{i0} = \left(\frac{\rho_d^{(N-i)}\left(1-(\rho_d/(N-i))^{(M-N+i+1)}\right)}{\left(1-\dfrac{\rho_d}{N-i}\right)(N-i)!} + \sum_{k=0}^{N-i-1} \rho_d^k / k! \right)^{-1}$$

$$i = 0,\cdots,N_v$$

The above represents a closed-form approximation to the GSM/GPRS network, for non-preemptive priority with single-slot GPRS data assignment.

The above approach was initially applied to GPRS in [1] based upon the decomposition method of [18]. The technique was additionally used in [3] to propose and analyze two alternate quick download mechanisms to support GPRS data users to receive the whole packet in time before moving of the cell. In [4] the technique is use to analyze GSM/GPRS networks in terms of the mean delay and 95% delay, and the impact of buffer assignment and guard channels on the GPRS traffic is also investigated.

A similar analysis can be carried out for more complex variations of the above scenario, such as for multi-slot GPRS, where GPRS data may use more than one channel or non-preemptive GSM voice priority over GPRS data. The latter case has been analyzed in [9], where it has been shown that the decomposition approximation can be usefully applied as well. However in this case a closed-form solution is not obtained, rather an iterative solution is used to find the solution.

IV. Isochronous and Asynchronous traffic modeling using self-similar processes

Although Poisson processes with exponential interarrival distributions have been frequently used in the analysis of networks due to their attractive theoretical properties, recent studies ([21], [22], [23]) have shown that traffic from a variety of packet networks at the (LANs, WANs, etc.) is much better modeled using statistically self-similar processes. Self-similar processes are characterized as having similar characteristics (such as burstiness) over a wide range of timescales. They also exhibit correlations over a wide range of timescales and they have long-range dependence. Such self-similar processes have quite different characteristics than Poisson processes.

It has been in shown in [21] that the superposition of many alternating independent and identically distributed ON/OFF sources whose ON and OFF periods have high variability or infinite variance results in self-similar aggregate traffic [24]. The Pareto distribution has been frequently used for the distribution of the ON period.

The Pareto distribution is defined as:

$$P(T \le t) = 1 - \left(\frac{k}{t}\right)^{\alpha} \qquad k, \alpha \ge 0, \qquad t \ge k$$

k is generally referred to as the location parameter, and α is the shape parameter. It has the property that if $\alpha \le 2$ the distribution has infinite variance, while if $\alpha \le 1$ the mean is infinite as well. The above definition results in a distribution that is heavy-tailed with infinite mean and variance.

The Decomposition approximation has been applied in [20], [6] to the analysis of various types of multiplexers that integrate Isochronous (guaranteed bandwidth) and Asynchronous (variable bit rate) such as ATM traffic.

The objective here is generally to have a control structure that minimizes a global cost function to maintain the Quality of Service (QoS) requirements of different users and service classes. For the isochronous traffic generally the call blocking probability needs to be minimized, and for the asynchronous traffic the cell loss probability and/or delay needs to be minimized.

In [20] control of a TDM Multiplexer has been analyzed, considering isochronous (circuit-switched) and asynchronous (packet-switched) traffic. The isochronous traffic is, in turn, subdivided into several classes, which are distinguished according to their speed, in order to model a multirate, multitraffic environment. Each class of isochronous traffic has been modeled as Poisson arrivals, with exponential holding times. The asynchronous traffic has been modeled by a M/Pareto/1/K queue using the Decomposition approximation. The aim of the analyses has been to define a control scheme for the allocation of the output link bandwidth, in order to realize two objectives:

a) To minimize call blocking probability for the isochronous traffic, and packet loss probability,
b) To meet quality of service requirements for both traffic types as closely as possible.

Using the large difference in time scales between the isochronous flows and the asynchronous ones, the Decomposition approximation has been used to find the stationary distribution of the M/Pareto/1/K queue model for the asynchronous traffic.

In [6] a some what similar analysis has been carried out in the ATM context, with a mix of Continuous Bit Rate (CBR) and Variable Bit Rate (VBR) traffic sharing a common bandwidth with Available Bit Rate (ABR) or Unspecified Bit Rate (UBR). A hierarchical decomposition has been used between the call-level and cell-level QoS. The network traffic is categorized into H+1 service classes, with the first H containing either CBR or bursty VBR (on-off) sources, characterized by parameters such as peak rate, average transmission rate and average burst length, as well as QoS requirements, like cell loss probability and cell delay. The asynchronous packet flow, which represents the traffic generated by connectionless, best effort services, is modeled as originating from the superposition of a number of on-off sources, whose duration follows a Pareto distribution. Here again, the Decomposition approximation has been used to decouple the CBR and VBR traffic.

In [8] an adaptive bandwidth allocation method in the Satellite environment has been analyzed. The paper proposes two alternative bandwidth allocation schemes suited for the Ka-band satellite environment. The objective here is to provide a control mechanism to compensate for rain fade. Both the schemes are organized in two hierarchical levels. The bandwidth allocation methods are aimed at keeping the call blocking probability of the guaranteed traffic below a given threshold and at reducing the packet dropping probability of the non-guaranteed, best-effort traffic. An ATM-based frame has been assumed, and the asynchronous traffic has been similarly modeled as in [6], and a similar Decomposition approximation has been used to simplify the traffic analysis.

However it is important to note here that the Decomposition approximation applied in all the above three cases [20], [6], and [8] is based on [18] where the approximation has only been verified for Poisson arrivals and exponentially distributed services times. Hence a future possible area of research is to validate the approximation for non-exponential distributions such as the self-similar Pareto distribution considered above.

V. Traffic modeling for use in Neural Network based admission control

In [5] and [7] neural networks have been used in the optimal admission control of an access node to a multiservice node, such as a base station in an integrated services cellular wireless network, or the optical line terminal (OLT) in a broad-band passive optical network (PON). The output link bandwidth is adaptively assigned to different users and dynamically shared between two types of traffics: isochronous (guaranteed bandwidth) and asynchronous as was the case in the applications discussed in the previous section. The optimal admission control strategies are approximated by means of backpropagation feedforward neural networks, acting on the embedded Markov chain of the connection dynamics. The neural networks operate in conjunction with a higher level bandwidth allocation controller, which performs a stochastic optimization algorithm. The goal of the admission controller is to minimize the refusal rate of connection requests as well as the loss probability of packets queued in a finite buffer.

The system model used in [5] is a TDM frame with a capacity of C slots. A free slot may be assigned to either a circuit-switched isochronous traffic or to packet-switched traffic with Poisson arrivals and fixed size packets that can be transmitted in a single slot (e.g. ATM cell). Use of the Decomposition approximation results in modeling the packet-switched traffic as an $M/D/(C-\tau)$ queue. Here τ is the number of isochronous calls in progress, each taking up one slot per frame.

In [7] a more complex model is used. An ATM access multiplexer is considered with a total capacity of C cells per a so called “frame”. The real-time isochronous traffic has preemptive priority over the asynchronous connectionless traffic. The isochronous traffic is assumed to have an exponential distribution for interarrival and holding times. The arrival process of the asynchronous traffic however is modeled as the superposition of

geometric distributions, a commonly used approach in modeling cell arrivals at ATM switches, with fixed length cell length. The queuing model is hence *Geom(N)/D/R/B* queue, with R servers available to N customers, each requesting service with probability p $(0 \le p \le 1)$ and B positions are available to hold the customers in the system. The system is synchronous in the sense that the time axis is slotted, each slot lasting one (deterministic) service time of a customer. All arrivals and departures take place at slot boundaries. The probability distribution of customer request A in a generic slot is binomial:

$$\Pr[A = i] = \binom{N}{i} p^i (1-p)^{N-i} \qquad (0 \le i \le N)$$

As in the previous section, a future area of research is to validate the Decomposition approximation for a geometric arrival distribution, as well as for fixed length packets.

VI. Conclusion and Future Research

In this paper we have reviewed various approximation methods that have been used in the performance evaluation of networks with various classes of traffic. In particular the Decomposition approximation has been studied and its applications in various recent papers have been reviewed. Applications in GSM/GPRS networks, TDM multiplexers, and ATM access multiplexers in optical and satellite networks have been considered. It has been highlighted that the original approximation considered a traffic model that was based on Poisson arrivals and exponentially distributed holding times for both the circuit-switched and packet-switched traffic. However the recent applications have applied the approximation to various distributions such as self-similar Pareto distributed processes, geometrically distributed arrival rates and fixed length packets transmitted as ATM cells. Hence a possible future area of research is to validate the approximation for the above scenarios. Additionally the approximation can be extended to the study of more complex scenarios such as multi-slot GPRS analysis, and ATM networks with multiple classes of traffic with possibly three or more time scales such as would be in the case of video, voice and data [19].

References

[1] Shaoji Ni, Sven-Gustav Häggman, "GPRS performance estimation in GSM circuit switched services and GPRS shared resource systems", *WCNC 1999 - IEEE Wireless Communications and Networking Conference,* no. 1, September 1999, pp. 1417-1421.

[2] Ying-Dar Lin, Yu-Ching Hsu, Mei-Yan Chiang, "Two-stage dynamic uplink channel and slot assignment for GPRS", *ICC 2001 - IEEE International Conference on Communications,* no. 1, June 2001, pp. 1335-1339.

[3] Haw-Yun Shin, Jean-Lien C. Wu, "The study of quick download mechanism in an infrastructural wireless environment", *GLOBECOM 2002 - IEEE Global Telecommunications Conference,* no. 1, November 2002, pp. 890-894.

[4] Hung-Huan Liu, Jean-Lien C. Wu, Wan-Chih Hsieh, "Delay analysis of integrated voice and data service for GPRS", *IEEE Communications Letters,* no. 8, Aug 2002, pp. 319-321.

[5] Bolla, R.; Davoli, F.; Maryni, P.; Parisini, T., "An adaptive neural network admission controller for dynamic bandwidth allocation", *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, Volume: 28, Issue: 4, Aug 1998, page(s): 592-601.

[6] Raffaele Bolla, Franco Davoli, Stefano Ricciardi, "A hierarchical control structure for multimedia access networks", *ICC 1999 - IEEE International Conference on Communications,* no. 1, June 1999, pp. 1320-1325.

[7] Franco Davoli, Piergiulio Maryni, "A two-level stochastic approximation for admission control and bandwidth allocation", *IEEE Journal on Selected Areas in Communications,* no. 2, February 2000, pp. 222-233.

[8] Raffaele Bolla, Franco Davoli, Mario Marchese, "Adaptive bandwidth allocation methods in the satellite environment", *ICC 2001 - IEEE International Conference on Communications,* no. 1, June 2001, pp. 3183-3190.

[9] S. Ghani and M. Schwartz, "A decomposition approximation for the Performance Evaluation of Non-Preemptive Priority in GSM/GPRS," submitted to *IEEE ICC 2004*, to be held on 20-24 June 2004.

[10] G. Coviello and P. A. Vena, "Integration of circuit/packet switching in a SENET (slotted envelope network) concept," *Nat. Telecommun. Conf. Rec.*, New Orleans, LA, Dec. 1975, pp. 42-12 – 42-17.

[11] K. Kummerle, "Multiplexor performance for integrated line- and Packet- Switched traffic," *ICCC*, Stockholm, Sweden, 1974, pp. 508-515.

[12] K. Sriram, P. K. Varshney, and J. G. Shanthikumar, "Discrete-time analysis of integrated voice/data multiplexers with and without speech activity detectors," *IEEE J. Select. Areas Commun.*, vol SAC-1, Dec. 1983, pp. 1124-1132.

[13] G. F. Williams and A. Leon-Garcia, "Performance analysis of integrated voice and data hybrid-switched links," *IEEE Trans. Commun.*, vol. COM-32, June 1984, pp. 695-706.

[14] P. O'Reilly, "Performance Analysis of Data in Burst Switching", *IEEE Transactions on Communications*, vol. Com-34, No. 12, December 1986, pp. 1259-1263.

[15] P. O'Reilly and S. Ghani, "Data Performance in Burst Switching when the Voice Silence Periods have a Hyperexponential Distribution", *IEEE Transactions On Communications*, vol. Com-35, No. 10, October 1987, pp. 1109 – 1112.

[16] D. P. Gaver and J. P. Lehoczky, "Channels that cooperatively service a data stream and voice messages," *IEEE Trans. Commun.*, vol. COM-30, May 1982, pp. 1153-1162.

[17] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis*. Reading, MA: Addision-Wesley, 1987.

[18] S. Ghani and M. Schwartz, "A decomposition approximation for the analysis of voice/data integration,"

in *IEEE Trans. Commun.*, vol. 42, pp. 2441-2452, July 1994.
[19] S. Ghani, "A Decomposition Approximation for Integrated Networks", PhD Dissertation, Columbia University, 1990.
[20] Raffaele Bolla, Franco Davoli, "Control of multirate synchronous streams in hybrid TDM access networks", *IEEE/ACM Transactions on Networking,* no. 2, Apr 1997, pp. 291-304.
[21] W. E. Leland, M. S . Taqqu, W. Willinger, and D. V. Wilson, "On the Self-similar Nature of Ethernet Traffic (Extended Version)", *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, February 1994.
[22] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling", *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, June 1995, pp. 226-244.
[23] V. Paxson, "Empirically derived analytic models of wide-area TCP connections", *IEEE/ACM Transactions on Networking*, No. 4, Aug 1994, pp. 316-336
[24] Z. Harpantidou, M. Paterakis, "Random Multiple Access of Broadcast Channels with Pareto Distributed Packet Interarrival Times", IEEE Personal Communications, April 1998, pp. 48-55.

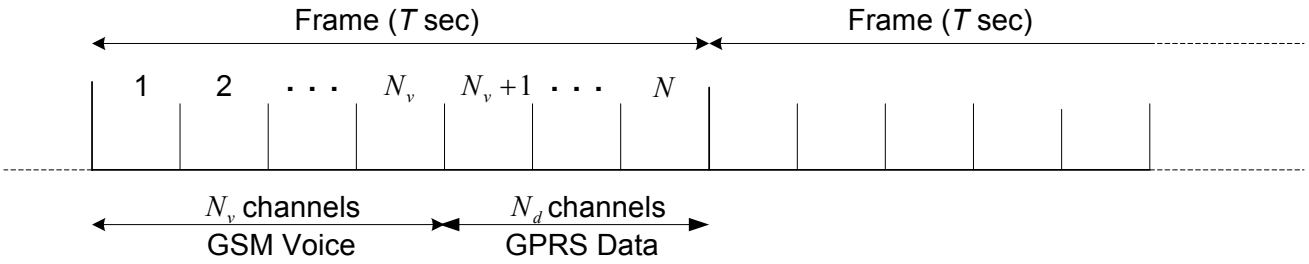

Figure 1. Frame Structure of GSM/GPRS

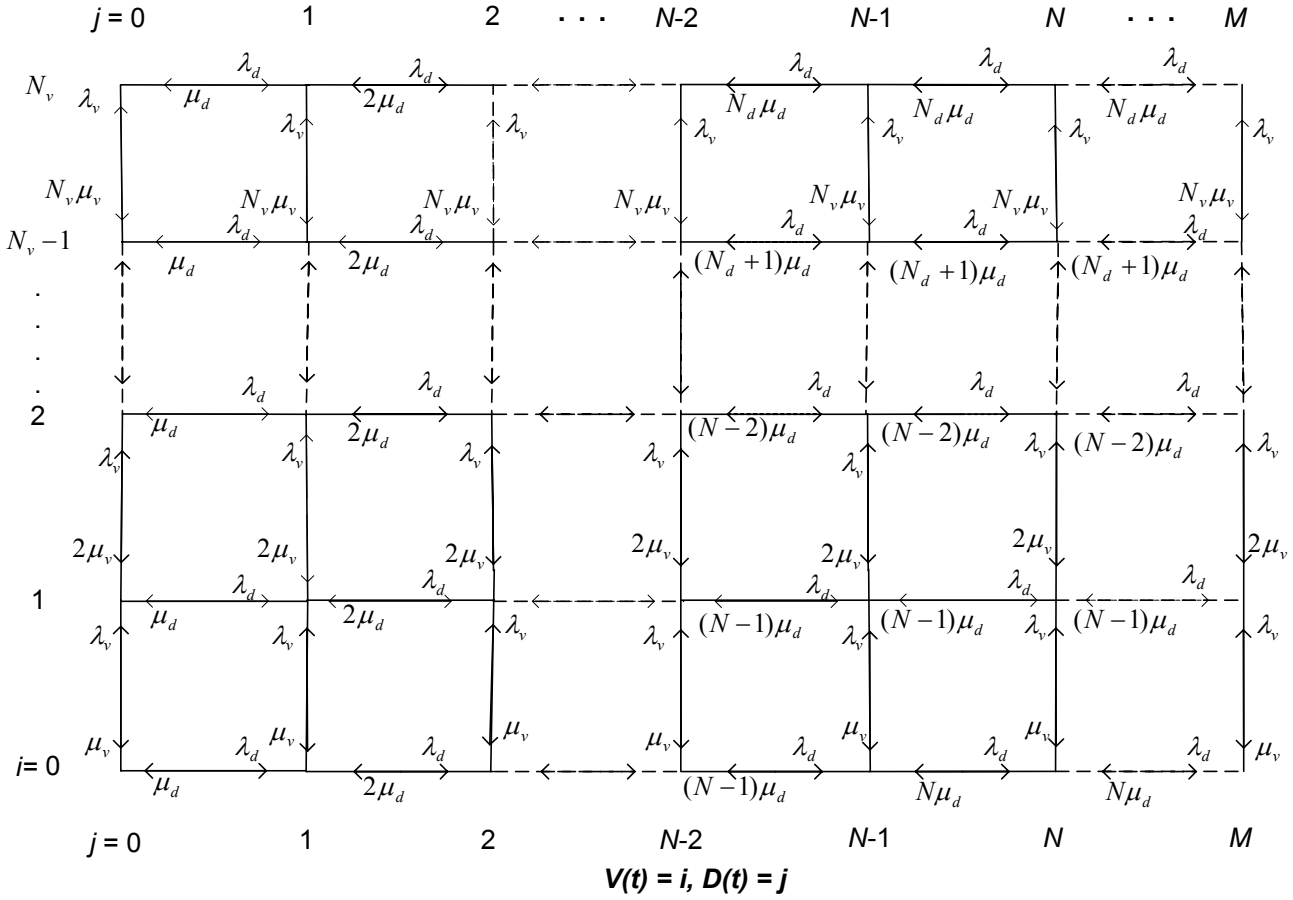

Figure 2. State Transition Diagram of GSM/GPRS