Performance Analysis of Communication Networks in Multi-Cluster Systems under Bursty Traffic with Communication Locality

Yulei Wu¹, Geyong Min¹, Keqiu Li², and Bahman Javadi³

¹School of Computing, Informatics and Media, University of Bradford, Bradford, BD7 1DP, UK
²Department of Computer Science and Engineering, Dalian University of Technology, P.R. China
³INRIA Rhone-Alpes, Laboratoire Informatique de Grenoble, ENSIMAG Grenoble, France
Email: {y.l.wu, g.min}@brad.ac.uk, keqiu@dlut.edu.cn, javadi@imag.fr

Abstract—Cluster-based systems have emerged as a promising technology for providing cost-effectiveness in high-performance computing and communication systems. Performance studies on communication networks in cluster-based systems have been reported based on the simplified assumptions that the traffic follows the non-bursty Poisson process and the message destinations are uniformly distributed over all network nodes. However, the uniform distribution of message destinations is not always realistic in practice. Moreover, the communication locality, a typical example of the non-uniform destination distribution, has been shown to be an important phenomenon in the communication networks of cluster systems. Many recent measurement studies have revealed that the traffic generated by many real-world applications exhibits a high degree of burstiness. In order to have a comprehensive understanding of the system performance, this paper proposes a new analytical model for communication networks in multi-cluster systems under the bursty message arrivals with communication locality. The model is validated through extensive simulation experiments.

1. INTRODUCTION

Cluster-based systems have emerged as a promising technology to provide cost-effectiveness in high-performance computing and communication systems for solving many complex scientific, engineering, and commercial applications. The communication network is an essential component in cluster-based systems for the achievement of high performance and the satisfaction of Quality-of-Service (QoS) requirements of diverse applications [3, 5]. Therefore, cluster-based systems must be constructed through interconnection that supports the full bandwidth to provide high performance and low latency communication infrastructure. Fat-trees have raised in popularity in cluster systems to provide such an infrastructure. The m-port n-tree [11], a typical example of the fat-tree topologies, has full bisectional bandwidth and can offer a rich connectivity among nodes, making possible to maintain paths between all source and destination nodes [7], which is critical for satisfying the performance requirements of cluster computing and communication systems. The m-port n-tree consists of \(2(m/2)^n\) processing nodes (PNs) and \((2n - 1)(m/2)^{n-1}\) communication switches (CSs) [11].

Message latency is the most important factor to test and evaluate the overall performance of communication networks in high-performance computing and communication environments [5]. It includes all possible delays encountered during the lifetime of a message. These delays can reflect the dynamic behaviours of communication networks.

Performance study of communication networks in cluster-based systems from exhaustive simulations would be highly time-consuming. Analytical modelling is a cost-effective tool to evaluate and optimise system performance under different design alternatives and various traffic conditions. Many analytical models for communication networks in cluster systems have been reported based on the simplified assumptions that the traffic arrivals follow the non-bursty Poisson process and message destinations are uniformly distributed over all network nodes [2, 14]. However, the assumption of uniform destination distribution is not always realistic in practice and the communication locality, a typical example of non-uniform destination distributions, has been shown to be an important phenomenon exhibited by many applications in cluster-based systems [1]. The communication locality exists when the mean inter-node distance is smaller than that in the uniform destination distribution [5]. As a result, each message consumes fewer resources and thus reduces contention. In addition, a number of measurement studies [4] have revealed that the traffic generated by many real-world applications exhibits a high degree of burstiness (i.e., time-varying arrival rates) and possesses correlations in the number of message arrivals. Therefore, the traditional models with the simplified assumptions cannot capture the characteristics of the realistic network traffic. Very recently, a performance model [9] was proposed to consider the phenomenon of communication locality in cluster systems with the message arrivals modelled by the Poisson process which cannot capture the bursty nature of message arrival process. To this end, this paper proposes a new analytical model to investigate the message latency in communication networks of multi-cluster systems based on m-port n-tree in the presence of communication locality with bursty message arrivals modelled by the well-known Markov-modulated Poisson process (MMPP). The MMPP is able to capture the properties of bursty message arrivals while it still remains analytically tractable [6]. The validity of the model is demonstrated by comparing analytical results to those obtained through extensive simulation experiments of the actual system.

The rest of this paper is organised as follows. Section 2 reviews some preliminaries. Section 3 derives the analytical
2. Preliminaries

2.1. Multi-Cluster System Architectures

This work studies the multi-cluster system consisting of $C$ clusters, each of which contains $N$ PNs associated with their own Local Memories (LMs), as depicted in Fig. 1. Each cluster has two communication networks: an IntrA-Communication Network (ACN) used for the purpose of message passing between nodes within the same cluster and an inter-Communication Network (ECN) used to transmit messages between clusters and manage the entire system. The clusters are connected to each other by an Integrated Communication Network (ICN). ECN and ICN are connected by a set of Concentrators/Dispatchers (C/D) [3], which are used to combine message traffic from/to one cluster to/from other clusters. Each communication network (i.e., ACN, ECN, and ICN) is constructed in the $m$-port $n$-tree topology.

Fig. 1 The multi-cluster system architecture

The construction of ACN and ECN based on $m$-port $n$-tree is straightforward. In ICN, each cluster can be viewed again as the PN of the tree topology. Let $(n + 1)$ denote the height of the tree in ICN. Since there are $C$ clusters in the system, $n_c$ can be obtained by solving the equation $C = 2^{(m/2)^n_c}$ as

$$n_c = \left\lfloor \frac{\log_2 C - 1}{\log_2 m - 1} \right\rfloor$$

2.2. Switching Methods & Routing Strategies

Wormhole switching has raised in popularity in cluster-based systems due to its low buffering requirements and it makes latency insensitive to the message distance in the absence of blocking [5]. In wormhole switching, a message is divided into a sequence of flits, each of a few bytes for transmission and flow control. When the header flit of a message arrives at a node, it must acquire a channel before being forwarded to the next node. The remaining flits of this message follow the header in a pipelined fashion. The tail of this message releases the channel as it passes. When the header is blocked due to contention for output channels or due to insufficient buffer space, remaining data flits are blocked in situ. In deterministic routing [5], a message traverses a fixed path between the source and destination, which simplifies the implementation, avoids message deadlock, and guarantees an in-order delivery. Very recently, Gomez et al. [7] have shown that deterministic routing can achieve a similar, and in some scenarios higher, level of performance than adaptive routing in fat-trees. Thus, we adopt a deterministic routing [9] based on Up*/Down* algorithm [15] to forward messages in the networks.

3. The Analytical Model

The model is based on the following assumptions, which are commonly used in the related studies [1, 2, 9, 12, 14].

a) The arrivals of messages generated by each source node are bursty and follow an independent two-state MMPP, denoted by $\text{MMPP}_s$, where the subscript $s$ represents the traffic generated by source nodes. The $\text{MMPP}_s$ is characterised by the infinitesimal generator $Q_s$ of the underlying Markov chain and rate matrix, $\Lambda_s$, as

$$Q_s = \begin{bmatrix} -\phi_{s1} & \phi_{s1} \\ \phi_{s2} & -\phi_{s2} \end{bmatrix} \quad \text{and} \quad \Lambda_s = \begin{bmatrix} \lambda_{s1} & 0 \\ 0 & \lambda_{s2} \end{bmatrix}$$

where the element $\phi_{s1}$ is the transition rate from state 1 to 2 and $\phi_{s2}$ is the rate out of state 2 to 1. $\lambda_{s1}$ and $\lambda_{s2}$ are the traffic rates when the Markov chain is in state 1 and 2, respectively.

b) Messages generated from each source node can be sent to ECN with the probability $\xi$ and to ACN with the probability $(1-\xi)$. Let us refer to the message sent to ECN and ACN as inter-cluster messages and intra-cluster messages, respectively. Since inter-cluster messages are directed to a node in any other clusters with equal probability, we define this probability as the degree of inter-cluster locality $\eta_{out}$, i.e., $\eta_{out} = \xi$ [9]. A message is directed to a node within the cluster with intra-cluster locality $\eta_{in}$.

c) Message length is fixed and is equal to $M$ flits. The length of each flit is $L_f$ bytes.

d) The network switches are input buffered and each channel is associated with a single flit buffer.

The mean message latency, $\overline{Lat}$, for the multi-cluster system can be obtained by the weighted sum of the message latency in intra-cluster communication networks (i.e., ACN), $\overline{Lat}_{in}$, and inter-cluster communication networks (i.e., ECN and ICN), $\overline{Lat}_{out}$, as

$$\overline{Lat} = (1-\xi)\overline{Lat}_{in} + \xi \overline{Lat}_{out}$$

In what follows, we will first determine the traffic characteristics on network channels in ACN, ECN, and ICN, respectively, and then derive the message latency in intra-cluster and inter-cluster communication networks, separately.

3.1. Traffic Characteristics on Network Channels in Each Communication Network

The traffic pattern, in the presence of communication locality, mainly affects the average number of channels that a message need cross to reach its destination, commonly known...
as mean message distance. In m-port n-tree, the probability of a message traversing 2j channels (j channels in ascending phase and j channels in descending phase) to reach its destination is \( P_j \). Different choices of \( P_j \) could lead to different distributions of message destinations and, consequently, different mean message distances. As described in Assumption (b), inter-cluster messages are destined to a node in any other clusters with equal probability. Since the number of PNs at distance \( 2n \) is \((m/2)^{n-1}(1-m)\) in m-port n-tree \([11]\), \( P_j \) can be given by

\[
P_j = \frac{1}{N-1} \left\{ \begin{array}{ll}
(m-1)(m/2)^{j-1} & j = n \\
(m/2-1)(m/2)^{j-1} & 1 \leq j < n
\end{array} \right.
\]

(4)

Intra-cluster messages are destined to a node in each cluster with the same locality, \( \eta_{in} \). We adopt the model proposed by Agrawal \([1]\) for the intra-cluster locality in our analysis. For a given source node, message destinations are chosen randomly from \( \eta_{in}N \) nodes centred at the source node. With this locality model, the probability of an intra-cluster message traversing 2j channels to reach its destination with the communication locality, \( \eta_{in} \), can be obtained based on the derivation of Eq. (4) as follows \([9]\):

\[
P_{j, \eta_{in}} = \frac{(m/2-1)(m/2)^{j-1}(m/2-1/\eta_{in})}{(m-1)(m/2)^{j-1}(m/2-1/\eta_{in})} \quad 1 \leq j < n
\]

(5)

The mean message distance for intra-cluster messages, \( \overline{d}_{in} \), and inter-cluster messages, \( \overline{d}_{out} \), can be given by

\[
\overline{d}_{in} = \sum_{j=1}^{n} 2j P_{j, \eta_{in}} \quad \text{and} \quad \overline{d}_{out} = \sum_{j=1}^{n} 2j P_j
\]

(6)

Recall that messages generated from each source node has the probability, \((1-\xi)\), to be sent to ACN with the mean message distance \( \overline{d}_{in} \), given by Eq. (6). Since there are \( N \) nodes in each cluster and the messages generated from these nodes are sent to 4nN channels in the network, the traffic arriving at a network channel in ACN is \( t_A \) times of the traffic generated by a source node. \( t_A \) can be expressed as

\[
t_A = \frac{N(1-\xi)}{4nN} \quad \overline{d}_{in}
\]

(7)

In addition, messages generated by each source node has the probability, \( \xi \), to be sent to ECN in the source cluster and through ICN to reach its destination node in ECN in the destination cluster with the mean message distance \( \overline{d}_{out}^{(E)} \) and \( \overline{d}_{out}^{(I)} \) in ECN and ICN, respectively. \( \overline{d}_{out}^{(E)} \) and \( \overline{d}_{out}^{(I)} \) can be obtained by Eq. (6). Similarly, the traffic arriving at a network channel in ECN and ICN is \( t_E \) and \( t_I \) times, respectively, of the traffic generated by a source node. Based on Eq. (7), \( t_E \) and \( t_I \) can be given by

\[
t_E = \frac{(N\xi + N\xi)\overline{d}_{out}^{(E)}}{4nN} \quad \text{and} \quad t_I = \frac{CN\xi\overline{d}_{out}^{(I)} + C}{4nC}
\]

(8)

where \( n_c \) is given by Eq. (1).

Since the splitting and superposition of multiple MMPPs give rise to a new MMPP \([6]\), let MMPP\(_{cA}^\text{cA} \), MMPP\(_{cE} \), and MMPP\(_{cI} \) denote the traffic arriving at a given network channel in ACN, ECN, and ICN, respectively. We will present the derivation of MMPP\(_{cA} \) as an example and the parameter matrices of MMPP\(_{cE} \) and MMPP\(_{cI} \) can be obtained accordingly. Generally, \( t_A \) is not an integer. We partition \( t_A \) into two parts, fractional part and integral part, denoted by \( F \) and \( Z \), respectively. Let MMPP\(_{F} \) represent the result of traffic from the splitting of MMPP\(_{c} \) with the splitting probability \( F \). According to the principle of splitting an MMPP \([6]\), the infinitesimal generator, \( Q_F \), and rate matrix, \( \Lambda_F \), of MMPP\(_{F} \) can be obtained as

\[
Q_F = Q_s \quad \text{and} \quad \Lambda_F = FA_s
\]

(9)

The traffic arrival rate and its covariance function play a critical role in the method for the superposition of multiple MMPPs \([8]\). MMPP\(_{cA} \) is chosen to match the following four statistical characteristics of the superposition of \( Z \) traffic flows modelled by MMPP\(_{cA} \) and one MMPP\(_{F} \): 1) mean, 2) variance, 3) third central moment, and 4) integral of the covariance function of the arrival rate. Based on the method presented in \([8]\), the infinitesimal generator, \( Q_{cA} \), and rate matrix, \( \Lambda_{cA} \), of MMPP\(_{cA} \) can be obtained.

3.2. Message Latency for Intra-Cluster Communication Networks

In this section, we will derive the message latency in intra-cluster communication networks (i.e., ACN), \( Lat_{in} \), which consists of three parts: a) mean network latency, \( T_{in} \), that is the time for a message to cross the network; b) the message waiting time at the source, \( W_{in} \); and c) the time for the tail to reach the destination, \( \tau_{in} \). Thus, \( Lat_{in} \) can be expressed as

\[
Lat_{in} = T_{in} + W_{in} + \tau_{in}
\]

(10)

The quantities \( T_{in} \), \( W_{in} \), and \( \tau_{in} \) will be derived in the subsequent subsections.

3.2.1. Mean Network Latency

Since an intra-cluster message may cross different number of channels to reach its destination, we take into account the network latency of a 2j-channel message (i.e., a message need traverse 2j channels to reach its destination) as \( T_j \). For the sake of illustration, we number the network stage as follows: the network stage numbering is based on the location of switches between the source and destination. The numbering starts from the stage next to the source (stage 0) and goes up as we get closer to the destination (stage \( K-1 \)). In m-port n-tree, the number of stages to be crossed for a 2j-channel message is \( K = 2j - 1 \). Since messages are transferred to the local nodes as soon as they arrive at their destinations, we first consider the message service time at the final stage and then continue the analysis backward to the first stage. The service time of a channel experienced by a 2j-channel message at stage \( K-1 \), \( T_{K-1,j} \), can be found as

\[
T_{K-1,j} = MT_{cn}^A
\]

(11)
where \( M \) is the message length in flits and \( t^A_{cm} \) is the time for a single flit to transmit on a PN-to-CS (or CS-to-PN) connection in ACN and can be computed as 
\[
t^A_{cs} = 0.5\theta^A_s + L_s \gamma^A_s ,
\]
where \( \theta^A_s \) and \( \gamma^A_s \) are the network latency and the transmission time of one byte (inverse of bandwidth) in ACN.

The service time at the internal stages \( k \) \( (0 \leq k \leq K-2) \) might be more, since a channel would be idle when the channels of subsequent stages are busy. Thus, the service time experienced by the 2-\( j \)-channel message, \( T_{k,j} \), on a channel at the internal stages can be found as
\[
T_{k,j} = M^A_{cs} + \sum_{i=k+1}^{K-1} W_{b_{i,j}}
\]
where \( t^A_{cs} \) is the time for a flit to transmit on a CS-to-CS connection in ACN and \( W_{b_{i,j}} \) is the blocking time that a channel acquires a channel at stage \( k \). \( t^A_{cs} \) can be given by 
\[
t^A_{cs} = \theta^A_s + L_s \gamma^A_s , \]
where \( \theta^A_s \) denotes the switch latency in ACN. \( W_{b_{i,j}} \) can be determined by the probability that a message is blocked at this stage, \( P_{b_{i,j}} \), and the waiting time experienced by the message to acquire a channel when blocking occurs, \( W_{c_{k,j}} \). Therefore, \( W_{b_{i,j}} \) can be expressed as
\[
W_{b_{i,j}} = P_{b_{i,j}} W_{c_{k,j}}
\]
To compute \( P_{b_{i,j}} \), let us first calculate the joint probability, \( P_{a_{i,b}} \) \( (0 \leq a \leq 1) \) and \( (b = 1, 2) \), that the channel is idle or busy and the underlying Markov chain of MMPP\(_{cs} \) is at the state \( b \). \( P_{a_{i,b}} \) can be determined using a bivariate Markov chain, where state \( \mathcal{S}_{a,b} \) corresponds to the case where the channel is idle \( (a = 0) \) or the channel is busy \( (a = 1) \), and the MMPP\(_{cs} \) is at state \( b \). The transition rate out of state \( \mathcal{S}_{a,b} \) to \( \mathcal{S}_{a+1,b} \) is \( \lambda^A_{a,b} \), where \( \lambda^A_{a,b} \) is the traffic rate on network channels in ACN. The service time at the internal stages can be found as
\[
T_{k,j} = M^A_{cs} + \sum_{i=k+1}^{K-1} W_{b_{i,j}}
\]
where the service time on network channels at stage \( k \) can be determined from the Laplace-Stieltjes transform (LST) of service time on network channels at stage \( k \) [10], denoted by \( T^A_{k,j} \). \( T^A_{k,j} \) will be derived in Section 3.2.2. The traffic intensity is \( \rho_{cs} = t^A_{cs} \lambda^A_{cs} \), where \( \lambda^A_{cs} \) is the mean traffic arrival rate at network channels in ACN and is equal to \( \pi^A_{cs} \lambda^A_{cs} \). \( \pi^A_{cs} \) is the steady-state vector of MMPP\(_{cs} \) and \( \lambda^A_{cs} = A_{cs} \hat{e}_{cs} \cdot e_{cs} \) is the column unit vector of length 2. The algorithm for computing the matrix \( e_{cs} \) can be found in [6].

Averaging over all possible nodes destined by a message in ACN, we can obtain the network latency as
\[
T_{in} = \sum_{j=1}^{n} P_{j,h_{in}} T_j
\]
where \( T_j \) is equal to the average service time of a message at stage \( 0 \), i.e., \( T_j = T_{0,j} \).

3.2.2 Laplace-Stieltjes Transform of the Service Time on Network Channels at Stage \( k \)

The calculation of \( t^A_{cs} \) and \( t^{(2)}_{cs} \) in Eq. (16) is driving the derivation of the LST of message service time on network channels in ACN, \( T^A_{k,j} \). Since the LST of the sum of independent random variables is equal to the product of their transforms [10], the LST of \( T_{k,j} \) can be expressed as
\[
T^A_{k,j} = \left\{ \begin{array}{ll}
\frac{1}{1 - e^{-K \mu_{cs}}} & k = K-1 \\
W^*_{k,j}(s) e^{-sK \mu_{cs}} & 0 \leq k \leq K-2
\end{array} \right.
\]
\[
W^*_{k,j}(s) = \sum_{a=0}^{K-1} \sum_{b=1}^{2} P_{a,b}
\]
wheel

Solving the above system of equations yields \( P_{a,b} \). Thus, the probability, \( P_{b_{i,j}} \), can be given by 
\[
P_{b_{i,j}} = \sum_{a=1}^{2} P_{a,b}
\]
To determine the waiting time, \( W_{c_{k,j}} \), the network channel is treated as an MMPP/G/1 queueing system [6]. As the arrival process is modelled by MMPP\(_{cs} \) and the service time is \( T_{k,j} \), \( W_{c_{k,j}} \) can be expressed as
\[
W_{c_{k,j}} = \frac{1}{2(1-\rho_{cs})}[2\rho_{cs} + \lambda^A_{cs} \gamma^A_{cs} - 2t^A_{cs}((1-\rho_{cs})g_{cs}]
\]
\[
+ t^A_{cs} (\pi^A_{cs} \hat{A}_{cs}) (Q^A_{cs} + \lambda^A_{cs} e_{cs} \pi^A_{cs})^{-1} \hat{A}_{cs}
\]
\[
W_{c_{k,j}} = \frac{2W_{c_{k,j}} - \lambda^A_{cs} \gamma^A_{cs}}{2\rho_{cs}}
\]
In the above equation, \( t^A_{cs} \) and \( t^{(2)}_{cs} \) denote the first two moments of the service time on network channels in ACN and can be determined from the Laplace-Stieltjes transform (LST) of service time on network channels at stage \( k \) [10], denoted by \( T^A_{k,j} \). \( T^A_{k,j} \) will be derived in Section 3.2.2. The traffic intensity is \( \rho_{cs} = t^A_{cs} \lambda^A_{cs} \), where \( \lambda^A_{cs} \) is the mean traffic arrival rate at network channels in ACN and is equal to \( \pi^A_{cs} \lambda^A_{cs} \cdot \pi^A_{cs} \) is the steady-state vector of MMPP\(_{cs} \) and \( \lambda^A_{cs} = A_{cs} \hat{e}_{cs} \cdot e_{cs} \) is the column unit vector of length 2. The algorithm for computing the matrix \( e_{cs} \) can be found in [6].

Averaging over all possible nodes destined by a message in ACN, we can obtain the network latency as
\[
T_{in} = \sum_{j=1}^{n} P_{j,h_{in}} T_j
\]
where \( T_j \) is equal to the average service time of a message at stage \( 0 \), i.e., \( T_j = T_{0,j} \).

3.2.3 Waiting Time at the Source

Recall that messages injected from a source node enter the ACN with the probability, \( (1-\xi) \). Thus, the traffic arriving at an injection channel in ACN, denoted by MMPP\(_{inf} \), is a
fraction of that generated by a source node. The fraction, $t_{sd}$, can be expressed as

$$t_{sd} = 1 - \xi$$ \quad (21)$$

Based on Eq. (9), the infinitesimal generator, $Q_{sd}$, and rate matrix, $A_{sd}$, of MMPP$_{sd}$ can be obtained. To determine the waiting time, $W_{in}$, that a message experiences before entering the network, the injection channel at the source node is modelled by an MMPP/G/1 queueing system, where the arrival process is modelled by MMPP$_{sd}$ and the service time is the network latency for an intra-cluster message given by Eq. (17). $W_{in}$ can be obtained in a similar manner to that used for the calculation of the message waiting time, $W_{ck}$, at a given network channel in ACN.

The mean time for the tail flit to reach its destination, $\tau_{in}$, in intra-cluster networks can be simply found as

$$\tau_{in} = \sum_{j=1}^{n} P_{j,\eta j} \left( \sum_{k=1}^{K-1} t_{ck} + t_{cn}^A + t_{cd}^A \right)$$ \quad (22)

### 3.3. Message Latency for Inter-Cluster Communication Networks

Inter-cluster messages traverse $2p$ channels in ECN ($p$ channels in the source cluster and $p$ channels in the destination cluster) and $2q$ channels in ICN to reach their destinations. Similarly to the intra-cluster messages, we take into account the network latency of a $2(p+q)$-channel message in inter-cluster networks (i.e., ECN and ICN) as $T_{p+q}$. The number of stages to be crossed for such a message is $K = 2(p+q) - 1$.

The service time of a channel experienced by a message at stage $k$, $T_{k,p+q}$, in inter-cluster networks can be obtained similarly to that for intra-cluster networks. Therefore, $T_{k,p+q}$ can be obtained as

$$T_{k,p+q} = \begin{cases} \sum_{i=k}^{K-1} W_{b,i,p+q} + M_{ek}^F, & 0 \leq k \leq K - 2 \\ M_{ek}^E, & k = K - 1 \end{cases}$$ \quad (23)

where $W_{b,k,p+q}$ is the blocking time that a message acquires a channel at stage $k$ in inter-cluster networks, and $t_{cn}^E$ is the time to transmit on a PN-to-CS (or CS-to-PN) connection in ECN and can be obtained based on the calculation of $t_{cn}^A$. $W_{b,k,p+q}$ can be calculated similarly to that used in the computation of $W_{b,k,i}$ in Eqs. (13)-(16). $t_{cs}$ can be found according to the transmission time of each flit in the correspondence networks.

Similarly to intra-cluster networks, the network latency experienced by a message in inter-cluster communication networks, $T_{out}$, can be calculated by

$$T_{out} = \sum_{p=1}^{n} \sum_{q=1}^{n_q} P_{p,q} T_{p+q}$$ \quad (24)

where $T_{p+q}$ is equal to the average service time of an inter-cluster message at stage 0, i.e., $T_{p+q} = I_0,p_q$; $P_{p,q}$ is the probability of a $2(p+q)$-channel message and $P_{p,q} = P_{p} P_{q}$.

Each inter-cluster message crosses a Concentrator (C) and a Dispatcher (D) during its network journey. The C/D is used to combine traffic from/to one cluster to/from other clusters and is working as a simple buffer [3]. To calculate the waiting time at this buffer, $W_{cd}$, we model the C/D as an MMPP/G/1 queueing system. The arrival traffic injected into this queue is MMPP$_{cd}$ and the service time is $M_{cd}^F$. $W_{cd}$ can be computed according to Eq. (16).

The message waiting time at the source in inter-cluster networks, $W_{out}$, can be computed in a similar manner to the calculation of that in the intra-cluster networks. Therefore, the message latency experienced in inter-cluster communication networks can be computed as follows:

$$Lat_{out} = T_{out} + W_{out} + 2W_{cd} + \tau_{out}$$ \quad (25)

where $\tau_{out}$ is the average time for the tail flit to reach its destination in inter-cluster communication networks and can be derived based on Eq. (22).

### 4. Validation of the Model

We have developed a simulator in OMNeT++ [13] to validate the accuracy of the analytical model. Extensive simulation experiments have been performed to validate the model. However, for the sake of specific illustration, latency results are presented for the following cases: number of clusters: $C = 2^4$; network topology: 4-port 3-tree; message length: $M = 64$ and 128 flits; flit length: $L_f = 512$ and 1024 bytes; the configuration of ACN is bandwidth is 800, network latency is 0.02, and switch latency is 0.01, and the configurations of ECN and ICN are bandwidth is 600, network latency is 0.1, and switch latency is 0.05; the intra-cluster locality, $\eta_s$, inter-cluster locality, $\eta_{out}$, and the infinitesimal generator, $Q_{s}$, of MMPP$_{s}$ are set in the captions of figures.

Fig. 2 depicts the results predicted by the model plotted against those obtained from the simulator as a function of the traffic rate in multi-cluster systems. In these figures, the horizontal axis represents the traffic rate, $\lambda$, at which a node injects messages into the network when the MMPP$_s$ is at state 1, while the vertical axis denotes the mean message latency obtained from the above model. For the sake of clarity of the figures, we have deliberately set the arrival rate, $\lambda$, at state 2 as zero; otherwise we need to use three-dimensional graphs to represent the results. These figures reveal that the message latency obtained from the above model closely match those obtained from the simulation.

### 5. Conclusions

This paper has proposed a new analytical model for the prediction of message latency in communication networks of multi-cluster systems in the presence of bursty traffic with communication locality. Each communication network in the proposed system has been constructed in the $m$-port $n$-tree. Extensive simulation experiments have been conducted to validate the accuracy of the analytical model. The tractability and accuracy of the analytical model make it a cost-effective tool to gain insight into the behaviour of communication networks in multi-cluster systems in the presence of bursty traffic with communication locality.
Fig. 2 Latency predicted by the model and simulation in 16 cluster systems with $\eta_{\text{in}} = 0.6$, $\eta_{\text{out}} = 0.5$: (a) $M = 64$, $\varphi_1 = 0.07$, $\varphi_2 = 0.07$, (b) $M = 128$, $\varphi_1 = 0.008$, $\varphi_2 = 0.008$.

REFERENCES


