time is the service time remaining for the frame in service upon i's arrival. Similarly, v_i is the residual vacation time seen by frame i upon arrival. Then:

$$w_i = r_i + v_i + \sum_{j=i-q_i}^{i-1} x_j + \left\lfloor \left(r_i + \sum_{j=i-q_i}^{i-1} x_j \right) \Big/ U \right\rfloor V \tag{7.16}$$

That is, the waiting time for frame i upon arrival is equal to the sum of four components: (1) residual service time for the frame in service, (2) the residual vacation time, (3) the time to service all frames in the queue, and (4) the vacations incurred while servicing the frames in front of the arriving one. This last component is vacation period V multiplied by the total number of cycles necessary to service the frame in service upon arrival plus the ones in the queue. The number of cycles in question is the floor of the total time to service the frames in the queue and the one in service upon arrival divided by the total operational time in each cycle.

The next step is to take the expectation of the terms in Equation 7.16. Nonetheless, expectation of the last term of the summation in the right side is difficult to compute. One can recur to the following approximation for low loads. From the point of view of the arriving frame, it is very likely that all work in front of it will be finished during one active period because the load in the system is small. In this case, $\left\lfloor \frac{1}{U}\left(r_i + \sum_{j=i-N_i}^{i-1} x_j \right) \right\rfloor = 0$.

The approximate expectation of the terms in Equation 7.16 the becomes:

$$E[w_i] = E[r_i] + E[v_i] + E[\sum_{j=i-q_i}^{i-1} E[x_j | q_i]] \tag{7.17}$$

$$E[w_i] = E[r_i] + E[v_i] + \frac{1}{\mu} E[q_i] \tag{7.18}$$

Taking the limit as $i \to \infty$:

$$E[w] = E[r] + E[v] + \frac{1}{\mu} E[q] \tag{7.19}$$

But by Little's Law [Kleinrock 75], $E[q] = \lambda E[w]$ and so:

$$E[w] = E[r] + E[v] + \frac{1}{\mu} \lambda E[w] \tag{7.20}$$

Finally:

$$E[w] = \frac{E[r] + E[v]}{1 - \rho} \tag{7.21}$$

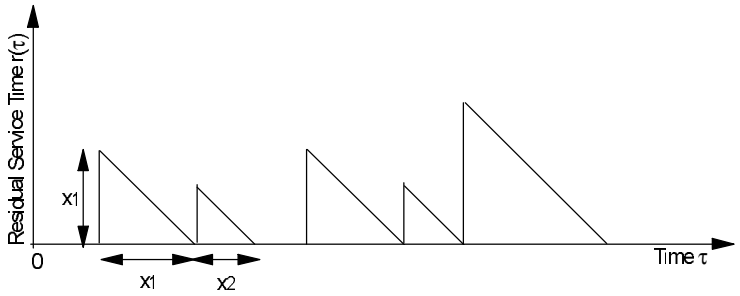

Figure 7.4: Residual service time

The mean residual service and vacation times can be computed using a graphical argument [Bertsekas and Gallager 92]. The residual service time is depicted in Figure 7.4. The residual time at *t*, *r*(*t*) is the time left to complete the frame transmission. When a new frame arrives, *r*(*t*) becomes equal to the service time for the frame. It then decreases linearly with time until it vanishes. The behavior is repeated for the next frame and so on.

The mean residual time at *t* can be computed from the time average of *r*(*t*), given that the number of frames serviced by time *t* is *M*(*t*), as:

$$\frac{1}{t} \int_0^t r(\tau) d\tau = \frac{1}{t} \sum_{i=1}^{M(t)} \frac{1}{2} x_i^2 \tag{7.22}$$

That is, the time average sum of the areas of all triangles in Figure 7.4 by time *t*. Equation 7.22 can also be written as:

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} x_i^2}{M(t)} \tag{7.23}$$

Taking the limits in Equation 7.23:

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\cdot\lim_{t\to\infty}\frac{\sum_{i=1}^{M(t)} x_i^2}{M(t)} \tag{7.24}$$

The first term in the right hand side multiplication is the one half of the input rate while the second is the expectation of the service time squared. The left hand side is the expectation of the residual service time. Formalizing this observation:

$$E[r] = \frac{1}{2}\lambda E[x^2] = \frac{1}{2}\lambda\overline{x^2} \tag{7.25}$$

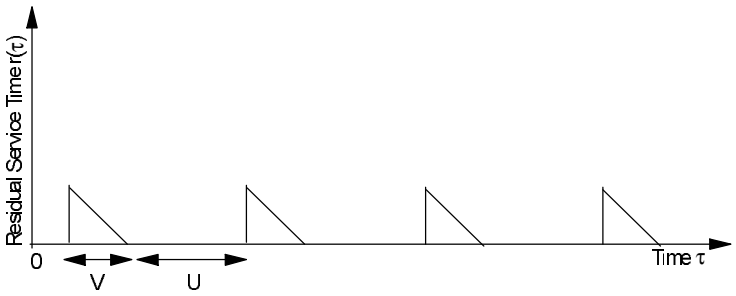

Figure 7.5: Residual vacation time

The mean residual vacation can be computed similarly to the mean residual service time. Figure 7.5 depicts the residual vacation time, where the interpretation is similar to the one for Figure 7.4. The average residual vacation is:

$$\frac{1}{t}\int_0^t v(\tau)d\tau = \frac{1}{t}\cdot\left\lfloor\frac{t}{V+U}\right\rfloor\cdot\frac{1}{2}V^2 \tag{7.26}$$

That is, the time average of the areas of the triangles in Figure 7.5 by time *t*. Taking the limits in Equation 7.28:

$$E[v] = \lim_{t\to\infty} \frac{1}{t}\int_0^t v(\tau)d\tau = \lim_{t\to\infty} \frac{1}{t}\cdot\frac{t}{V+U}\cdot\frac{1}{2}V^2 = \frac{1}{2}\frac{V^2}{V+U} \tag{7.27}$$

Replacing Equation 7.25 and 7.27 in Equation 7.21:

$$E[w] = \frac{1}{2}\frac{\lambda\overline{x^2}}{(1-\rho)} + \frac{1}{2}\frac{V^2}{(U+V)}\frac{1}{(1-\rho)} \tag{7.28}$$

From which it can be concluded that:

$$D = \frac{1}{2}\frac{\lambda\overline{x^2}}{(1-\rho)} + \frac{1}{2}\frac{V^2}{(U+V)}\frac{1}{(1-\rho)} + \frac{1}{\mu} \tag{7.29}$$

Notice that Equation 7.28 can be interpreted as follows. The first component of the summation is the normal waiting time of an M/G/1 system. The second component represents the penalty due to the residual vacation time. It is equal to the mean vacation period ($\frac{V}{2}$) times the probability of being on vacation ($\frac{V}{U+V}$) divided by the mean idle period ($1-\rho$) to obtain the total waiting time due to residual vacations.

Finally, there is no loss in RDMA++ and thus:

$$L = 0 \tag{7.30}$$

An interesting conclusion that can be derived from the RDMA++ analysis is that the same approximation for mean delay can be used in CS. CS operates similarly to RDMA++. The circuit comes periodically within a cycle, in similarity to a band arriving within a cycle. In conclusion, Equations 7.29 and 7.30 are also valid approximations for CS where *U* should be interpreted as the circuit duration.

7.2.3 The RDMA+ System

If the band size U is large compared to the mean transmission rate μ, then one can assume that RDMA+ operates in similarity to a M/G/1 queueing system with periodic losses. Figure 7.6 illustrates the typical behavior of an RDMA+ system. Frames behave like in RDMA++, except when the band ends. At that point, all frames in the system are lost. For example, in the figure, frame C_{n+1} is lost.

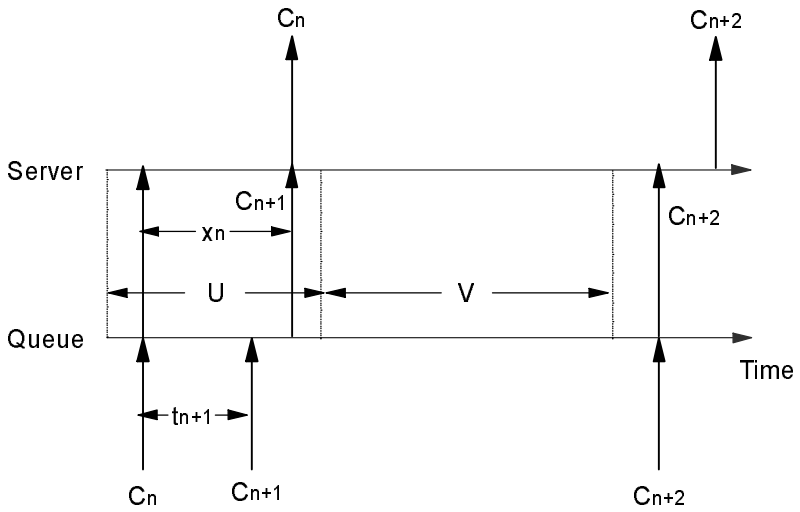

Figure 7.6: RDMA+ model

The model adopted here is just an M/G/1 queueing system. Delays are computed directly using M/G/1 delay formulas [Kleinrock 75]. For example, the mean delay is:

$$D = \frac{\lambda \overline{x^2}}{2(1-\rho)} + \frac{1}{\mu} \tag{7.31}$$

The delay distribution can be computed similarly using the M/G/1 formulas.

The loss probability can be approximated observing the following. If the band is big enough such that the system eventually becomes stable, then the probability that n frames will be lost at the end of the band is equal to the probability that n frames are in the system

at any time. This is the distribution of the number in the system for an M/G/1 system. The average number in the system is:

$$\overline{N} = \frac{\lambda^2 \overline{x^2}}{2(1-\rho)} + \rho \tag{7.32}$$

From which it can be concluded that the average loss in a band of size U is:

$$L = \frac{\frac{\lambda^2 \overline{x^2}}{2(1-\rho)} + \rho}{\lambda U} = \frac{\lambda \overline{x^2}}{2U(1-\rho)} + \frac{1}{\mu U} \tag{7.33}$$

It is interesting to notice in RDMA+ that when $U \to \infty,\ L \to 0$.

7.2.4 Network of Nodes

Somewhat surprisingly, the results in the previous sections can be used to approximate the behavior in a network of nodes. This section gives some insight on why the approximations are also valid for a network.

In RDMA, the dynamics in each routing tree are independently of the ones in other routing trees. For this reason, one can take each tree in isolation and analyze it. Additionally, each tree can be seen as a server with multiple incoming links, as in Figure 7.1. There are only a few reasons why both models, the tree model and the node model, are not equivalent. All of them have to do with the number of servers and links that each frame crosses and with the order in which frames are serviced.

The first is that the order of service can vary. In the tree model, sources that are close to the destination are serviced before sources that are distant from the destination. Nevertheless, the order of service is not important to compute both the mean delay and loss, so this difference can be ignored.

The second is that the number of servers that each frame crosses varies. This can be accounted for by analyzing each outgoing link in each tree separately. Then, the results (for delay and loss) may be combined using proper weights (based on the load of each link) to find the end-to-end delay.

The third is that the total propagation delay per frame may also vary according to how many links it crosses. This can be accommodated similarly by computing the average propagation delay taking into account the percentage of frames that will incur the given delay.

Finally, if the service distribution is not Poisson, traffic being input from a node will not present a Poisson distribution when input to the next node. One can approximate the result by assuming that the input to each link is indeed Poisson at each stage of the tree.

One can observe that these are common practices in analyzing network. For example, Reference [Kleinrock 76] makes similar assumptions and can be used to understand how the analysis on a single node can be extended to a network of nodes using the guidelines presented in this section.

7.3 Simulation: Contention for Outgoing Links

The goal of this section is twofold. Firstly, it validates the analysis in Section 7.1 with the simulation of the system depicted in Figure 7.1. Secondly, it compares the behavior of RDMA with PS and CS in a node.

The physical system studied is the one in Figure 7.1 with the following features. There are 10 input links to the node and 1 outgoing link that is shared among the inputs. Each link operates either at 2.4 Gb/s or at 24 Gb/s and transmits ATM cells (each 53 bytes in

size). The arrival process is Poisson. The frame size is either deterministic or distributed according to an exponential distribution. For the RDMA simulations, the cycle period is 125 μs. The band size varies in each case. Finally, the network topology is such that there are 5 bands in each cycle.

7.3.1 RDMA++

The purpose of the study is to show how closely the approximated and simulated RDMA++ mean frame delays match each other. As expected, the approximation works best for low to medium network loads in all experiments conducted.

Figure 7.7 illustrates mean delay when the frame size is fixed at 53 bytes, and the maximum link rate is 2.4 Gb/s. For this figure and the following, the study varies the input load up to 90% of its maximum capacity to avoid unstable behaviors as $\rho \to \frac{U}{U+V}$. The input load axis is measures in thousands of frames per second (Kframes). For example, at 73.584906 Kframes/s, the nominal input rate is 31.2 Mb/s in each input link. Because there are 10 input links, the aggregate input to the outgoing link is 312 Mb/s or 13% of the maximum rate of 2.4 Gb/s. Notice that, because the band size is 20% of the cycle, the maximum aggregate input rate to the outgoing link must be less than 20% or else the system will be unstable.

In Figure 7.8, the maximum link rate is 2.4 Gb/s, and the frame size is exponentially distributed with mean 53 bytes.

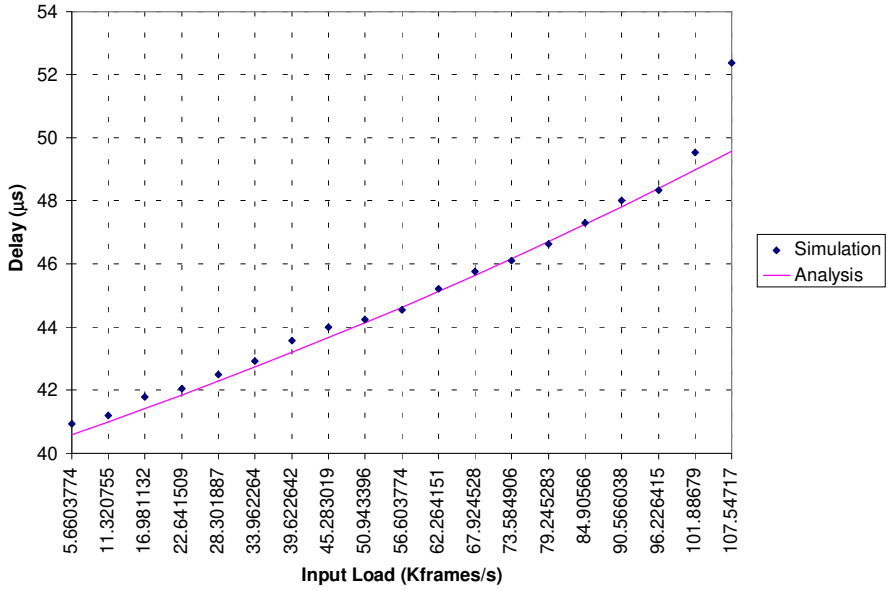

Figure 7.7: Mean frame delay (in µs) for Poisson arrivals and deterministic service under RDMA++ (the frame size is 53 bytes, the link speed 2.4 Gb/s, the band size 25 µs, the cycle period 125 µs)

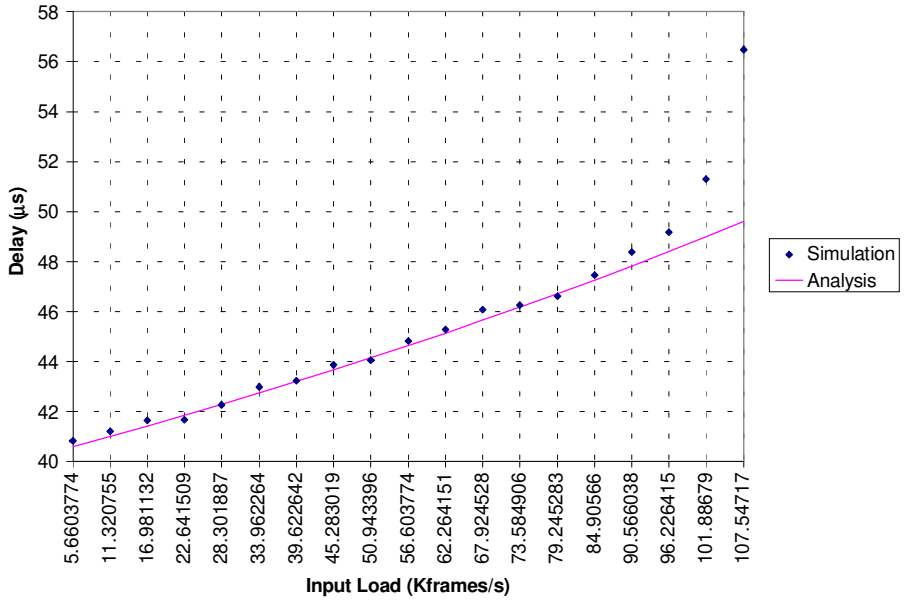

Figure 7.8: Mean frame delay (in µs) for Poisson arrivals and Poisson service under RDMA++ (the mean frame size is 53 bytes, the link speed 2.4 Gb/s, the band size 25 µs, the cycle period 125 µs)

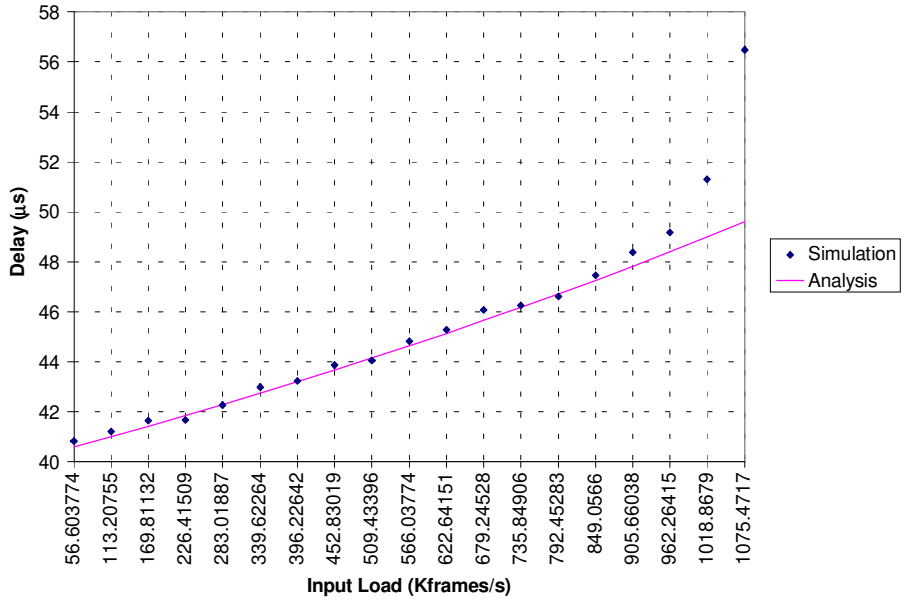

Figure 7.9: Mean frame delay (in µs) for Poisson arrivals and Poisson service under RDMA++ (the mean frame size is 530 bytes, the link speed 24 Gb/s, the band size 25 µs, the cycle period 125 µs)

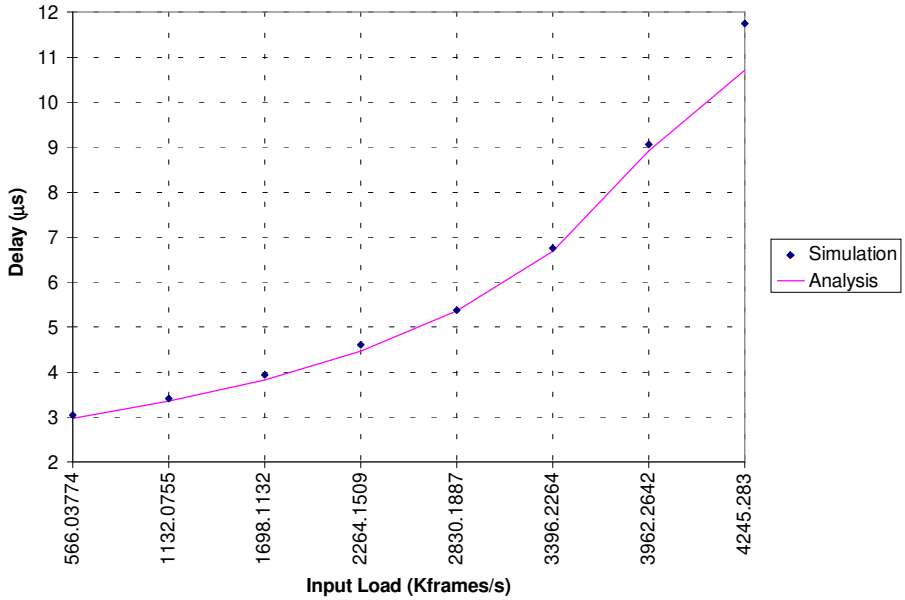

Figure 7.10: Mean frame delay (in µs) for Poisson arrivals and deterministic service under RDMA++ (the frame size is 530 bytes, the link speed 24 Gb/s, the band size 100 µs, the cycle period 125 µs)