BASIC MODELS OF CLOSED QUEUING NETWORKS

Dr. Jozo Dujmović

We present a summary of two important models of closed queuing networks. The first is the model of asymptotic behavior of interactive and batch systems. The second model is the mean value analysis that can be used

Asymptotic Analysis of Interactive and Batch Systems

An analysis of asymptotic behavior of closed queuing networks can be made assuming the model of *N* customers and *K* service stations presented in Fig. 1. Each service station is a queue and a server. This model can be used for the analysis of interactive transaction processing systems. However, in a special case where there are no interactive workstations We assume that each interaction (or each batch job) may cause multiple visits to each service station. The asymptotic behavior corresponds to the extreme cases of the very small and the very large number of customers (*N*=1 and $N \gg 1$)

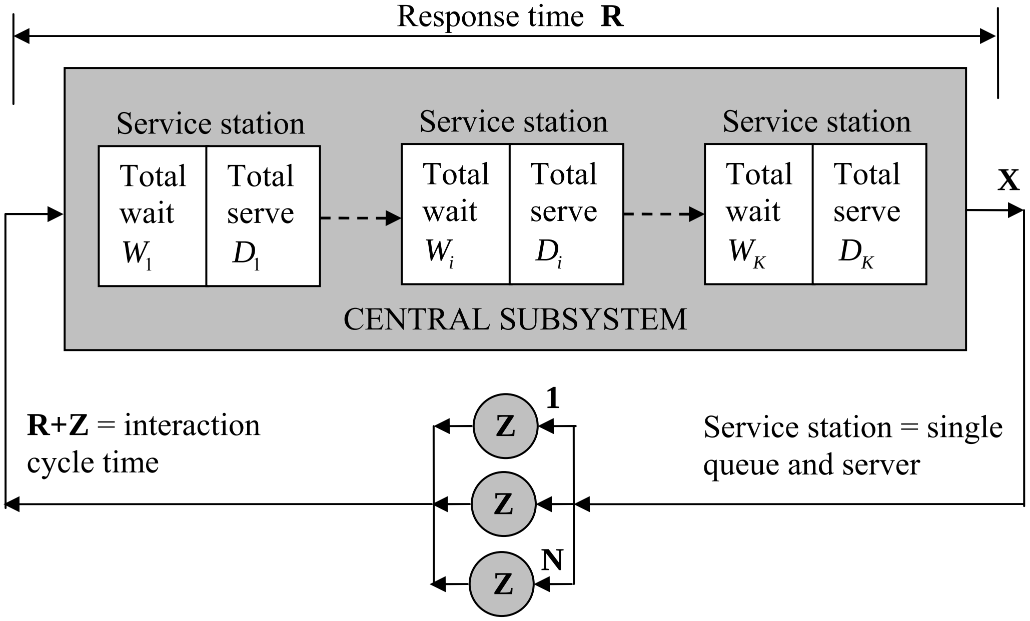


Figure 1. Workstations and the central subsystem (for batch processing models Z=0)

The model if this system includes the following variables:

N = number of interactive workstations
Z = mean think time (for batch models Z=0)
K = number of service centers in the central subsystem

$S_1,...,S_K$ = service times in K service centers

$V_1,...,V_K$ = number of visits to K service centers (per interactive transaction or per batch job)

$D_1,...,D_K$ = total demands for resources in K service centers per transaction or per job

$D_i = V_i S_i, \quad i = 1,...,K$

$X_1,..., X_K$ = throughput of K service centers (service requests per second)

X = throughput of central subsystem (transactions or jobs per second)

$U_1,...,U_K$ = utilization of servers in K service centers ($U_i = S_i X_i, \quad i = 1,...,K$)

$R(N)$ = average response time in the case of N workstations

$R(N)+Z$ = interaction cycle time (time between the beginnings of two successive interactions)

Dynamic behavior of the queuing network shown in Fig. 1 is described by the following models.

Forced flow law: the throughput of each resource equals the system throughput multiplied by the number of visits to that resource:

$$X_i = V_i X, \quad i = 1,...,K$$

Bottleneck device: the device that has the highest demand (denoted with index *b*):

$$D_b = D_{\max} = \max(D_1,...,D_K)$$

Maximum system throughput: the reciprocal of the maximum demand, $X_{\max} = 1/D_{\max}$:

$$U_b = S_b X_b = S_b V_b X = D_b X = D_{\max} X \le 1$$
$$X \le 1/D_{\max} = X_{\max}$$

Throughput law. During each interaction cycle time the central subsystem must serve all *N* customers (the quality of service must be the same for all customers). Therefore, the system throughput is

$$X(N) = \frac{\text{the number of customers}}{\text{interaction cycle time}} = \frac{N}{R(N)+Z}$$

Response time formula. It directly follows from the throughput law:

$$R(N) = \frac{N}{X(N)} - Z$$

For batch systems Z=0 and the response time formula reduce to the Little law:

$$R(N) = \frac{N}{X(N)}$$

Minimum response time for N=1. If N=1 there is no waiting in queues and all individual service center response times reduce to service times. Consequently, the system response time is the sum of all demands:

$$R(1) = D = D_1 + ... + D_K = R_{\min}$$

Response time for N>>1. If N>>1 then the throughput achieves the maximum value $X(N) \to X_{\max}$ and the response time approaches the following asymptote:

$$R(N) \to \frac{N}{X_{\max}} - Z = ND_{\max} - Z$$

Asymptotes of response time. The response time R(N) has two asymptotes: $R = D = D_1 + ... + D_K$ and $R = ND_{\max} - Z$ shown in Fig. 2. Here D denotes the sum of all demands which is the minimum response time. (Strictly speaking $R = D$ is not an asymptote because R(N) achieves this value for N=1.) Therefore, the general form of function R(N) is also shown in Fig. 2.

Asymptotes of throughput. The throughput law for N=1 yields

$$X(1) = \frac{1}{D+Z}$$

Therefore, the "asymptote" (in the same sense as for the response time) for throughput in the case of low load is a line going through the point (0,0) and (1, 1/(D+Z));

$$X(N) = \frac{N}{D+Z}$$

If N>>1 the real asymptote of X(N) is $X_{\max} = 1/D_{\max} = 1/D_b$. This asymptotes and the characteristic shape of X(N) are also shown in Fig. 2.

Critical number of workstations or jobs N*. The intersection of both the response time and the throughput asymptotes denotes the beginning of saturation phenomena caused by the bottleneck resource. Equations the determine the intersection are:

$$\frac{N^*}{D+Z} = \frac{1}{D_{\max}} \ , \qquad D = N^* D_{\max} - Z$$

The critical number of workstations and the critical number of batch jobs are

$$N^{*} = \frac{D+Z}{D_{\max}} = \frac{D_1 + \ldots + D_K + Z}{\max(D_1,\ldots,D_K)}, \qquad \left[N^{*} = \frac{D}{D_{\max}} = \frac{D_1 + \ldots + D_K}{\max(D_1,\ldots,D_K)}, \quad if \quad Z = 0 \right]$$

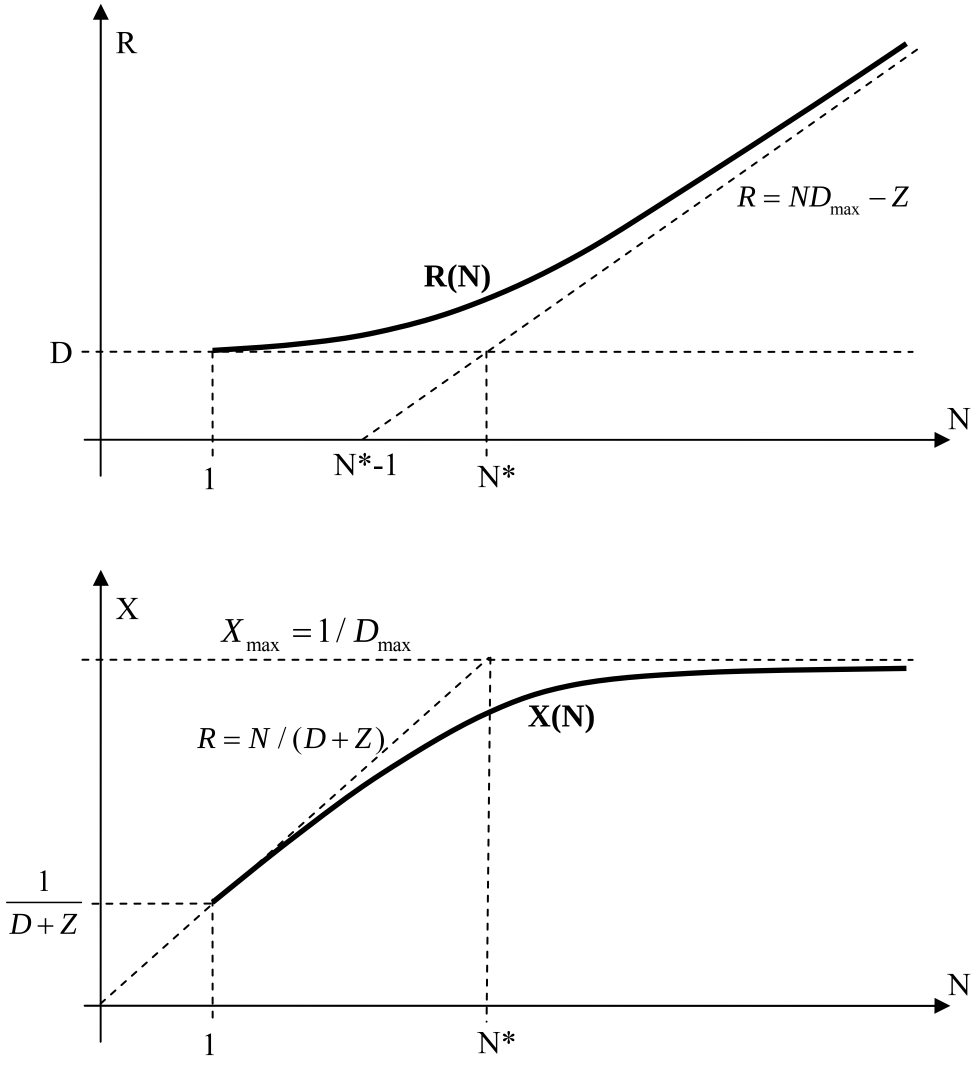


Figure 2. Asymptotes of the response time and throughput, and the critical number N*

Mean value analysis (MVA)

In order to analyze the dynamic behavior of closed queuing networks it is important to understand what happens to customer (job or transaction or service request) at the moment of arrival to the ith service center. This situation is presented in Fig. 3.

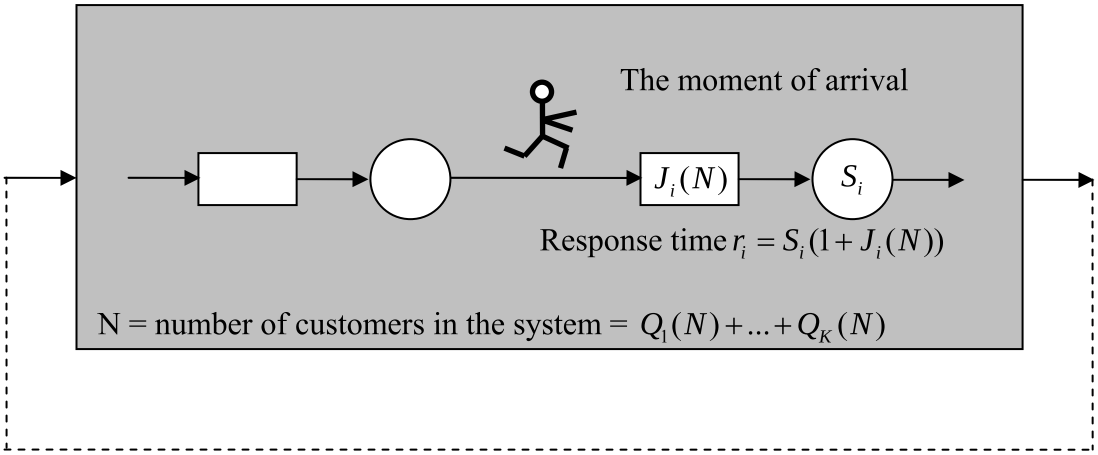

The moment of arrival
$J_i(N)$
S_i
Response time $r_i = S_i(1+J_i(N))$
N = number of customers in the system = $Q_1(N)+...+Q_K(N)$

Figure 3. The moment of transition of customer from a service center to another service center

Let us now focus on that specific point in time when a customer makes transition from a service center to another service center. The “moment of arrival” to the ith service center is characterized by $J_i(N)$ customers that the arriving customer encounters in the ith service center at the moment of arrival. That is a real number. For example, $J_i(N) = 3.5$ means that the arriving customer found 3 customers waiting in the queue and one customer that was receiving service, and at the moment of arrival already received 50% of the average service time S_i. Consequently, the time necessary to serve all encountered customers in the ith service center is $J_i(N)S_i$. This is also the time that the arriving customer must spend waiting in the queue before it can start receiving S_i units of service. Therefore, the time from the arrival to the departure of the customer, which is called the response time is

$$r_i(N) = J_i(N)S_i + S_i = S_i(1+J_i(N))$$

So, now we face the question how to determine $J_i(N)$. It is important to note that $J_i(N)$ is *not* the average number of customers that and external observer can measure over an infinite observation period. Indeed, over an infinite period of time *N* customers are distributed in *K* service centers so that $Q_i(N)$ customers is on the average in the ith service center, and consequently $Q_1(N)+...+Q_K(N) = N$. We know that $J_i(N) \neq Q_i(N)$. It is reasonable to expect that $J_i(N) < Q_i(N)$ because $J_i(N)$ reflects the moment when the customer is not in any of service centers and $Q_i(N)$ reflects the situation where the analyzed N$^{\text{th}}$ customer is present in service centers. In fact, the distributions of customers in service centers in the case of N-1 and N customers in the system are

$$\begin{aligned}&Q_1(N)+...+Q_K(N)=N\\&Q_1(N-1)+...+Q_K(N-1)=N-1\\&[Q_1(N)-Q_1(N-1)]+...+[Q_K(N)-Q_K(N-1)]=1\end{aligned}$$

Consequently, $Q_i(N)>Q_i(N-1)$ and $Q_i(N)-Q_i(N-1)$ can be interpreted as the probability that the N$^{\text{th}}$ customer is present in the i$^{\text{th}}$ service center.

It is possible to strictly prove the following fundamental MVA result:

$$J_i(N)=Q_i(N-1),\ i=1,...,K$$

This is what we might have expected: the arriving customer at the moment of arrival encounters the same number of customers that an external observer can measure in the case of N-1 customers in the system. That is sometimes interpreted by saying that "the arriving customer doesn't see himself" in the system because at the moment of transition the customer is in fact outside of system (not in any of service centers).

The presented MVA result yields the response time formula

$$r_i=S_i(1+J_i(N))=S_i(1+Q_i(N-1))$$

If we multiply this formula by the number of visits V_i the result will be the residence time R_i that determines the total time that a customer accumulates in the i$^{\text{th}}$ service center during all visits:

$$R_i=V_ir_i=V_iS_i(1+Q_i(N-1))=D_i(1+Q_i(N-1))$$

Of course, the response time is then the sum of all residence times:

$$R(N)=R_1(N)+...+R_K(N)$$

Initially, if N=0 (no customers in the system), all queues are empty:

$$Q_1(0)=...=Q_K(0)=0$$

In the case of *N* customers we can compute the throughput at each service center and the utilization of each server using the forced flow law as follows:

$$X_i = V_i X$$
$$U_i = S_i X_i = S_i V_i X = D_i X, \quad i = 1,...,K$$

If we want to analyze closed queuing networks with load independent servers (i.e. with service centers that consist of a single queue and a single server) then we can use the above formulas to create the following load independent MVA algorithm:

Input values:
Z = think time (in the special case of batch processing Z=0)
N = the number of workstations or jobs
K = the number of service centers
$D_1,...,D_K$ = service demands

```
for(i=1; i<K; i++) $Q_i$ =0                          // initial conditions
for(j=1; j<=N; j++)                                  //  j denotes the number of jobs
{
        for(i=1; i<K; i++) $R_i = D_i(1+Q_i)$        // residence times

        $R = R_1 + ... + R_K$                        // response time R
        $X = \frac{j}{Z+R}$                          // throughput for j jobs
        for(i=1; i<K; i++) $Q_i = XR_i$              // queue lengths for j jobs
}
$X_i = V_i X, \quad i = 1,...,K$
                                                     // throughputs and utilizations
$U_i = D_i X, \quad i = 1,...,K$
```

Results:
R = system response time
X = system throughput
$Q_1,...,Q_K$ = queue lengths at service centers
$X_1,...,X_K$ = throughputs at service centers
$U_1,...,U_K$ = server utilizations at service centers

This algorithm is valid both for interactive systems (where Z > 0) and for batch systems (where Z = 0). In analytic computation this algorithm can be conveniently implemented in a form of table (shown below).

PERFORMANCE ANALYSIS OF AN INTERACTIVE SYSTEM

Input data:

Maximum number of terminals =		Average think time [sec] =	
Resource identifier j	Resource type	Number of resources M_j	Total demand per transaction [sec] D_j
1	Central processor	1	
2	Disk channel		
3			
4			

Mean Value Analysis:

i	Residence time $R_j(i)=D_j(i)[1+Q_j(i-1)]$				Response time $R(i)$ $=\sum_j M_j R_j(i)$	Throughput $X(i)$ $= \frac{i}{Z+R(i)}$	Queue lenght $Q_j(i)=X(i)R_j(i)$			
	j=1	j=2	j=3	j=4			j=1	j=2	j=3	j=4
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										

PERFORMANCE ANALYSIS OF A BATCH SYSTEM

Input data:

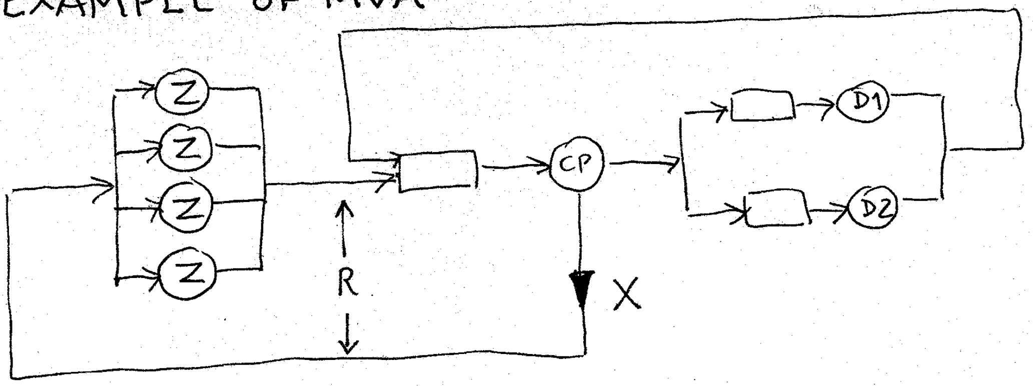
Maximum degree of multiprogramming =

Resource identifier **j**	Resource type	Number of resources $\mathbf{M}_j$	Total demand per transaction [sec] $\mathbf{D}_j$
1	Central processor	1	
2	Disk channel		
3			
4			

Mean Value Analysis:

i	Residence time $R_j(i)=D_j(i)[1+Q_j(i-1)]$				Response time $R(i)$ $=\sum_j M_j R_j(i)$	Throughput $X(i)$ $= \frac{i}{R(i)}$	Queue lenght $Q_j(i)=X(i)R_j(i)$			
	j=1	j=2	j=3	j=4			j=1	j=2	j=3	j=4
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										

AN EXAMPLE OF MVA



$N = 4$, $Z = 10$ sec

$D_p = 0.5$ sec

$D_1 = 2$ sec

$D_2 = 1$ sec

$Q_p(0) = 0$, $Q_1(0) = 0$, $Q_2(0) = 0$ Initial conditions

$m = 1$

$R_p(1) = D_p(1 + Q_p(0)) = D_p = 0.5$

$R_1(1) = D_1(1 + Q_1(0)) = D_1 = 2$

$R_2(1) = D_2(1 + Q_2(0)) = D_2 = 1$

$R(1) = R_p(1) + R_1(1) + R_2(1) = 3.5$ sec

$$X(1) = \frac{1}{Z + R(1)} = \frac{1}{10 + 3.5} = \frac{1}{13.5} = 0.074$$

$Q_p(1) = X(1)\, R_p(1) = 0.074 \cdot 0.5 = 0.037$

$Q_1(1) = X(1)\, R_1(1) = 0.074 \cdot 2 = 0.148$

$Q_2(1) = X(1)\, R_2(1) = 0.074 \cdot 1 = 0.074$

$n=2$

$R_p(2) = D_p\,(1+Q_p(1)) = 0.5 \cdot 1.037 = 0.5185$

$R_1(2) = D_1\,(1+Q_1(1)) = 2 \cdot 1.148 = 2.296$

$R_2(2) = D_2\,(1+Q_2(1)) = 1 \cdot 1.074 = 1.074$

$R(2) = R_p(2)+R_1(2)+R_2(2) = 3.8885$

$$X(2) = \frac{2}{Z+R(2)} = \frac{2}{13.8885} = 0.144$$

$Q_p(2) = X(2)\,R_p(2) = 0.144 \cdot 0.5185 = 0.07466$

$Q_1(2) = X(2)\,R_1(2) = 0.144 \cdot 2.296 = 0.3306$

$Q_2(2) = X(2)\,R_2(2) = 0.144 \cdot 1.074 = 0.1547$

$n=3$

$R_p(3) = D_p\,(1+Q_p(2)) = 0.5 \cdot 1.07466 = 0.53733$

$R_1(3) = D_1\,(1+Q_1(2)) = 2 \cdot 1.3306 = 2.6612$

$R_2(3) = D_2\,(1+Q_2(2)) = 1.1547$

$R(3) = R_p(3)+R_1(3)+R_2(3) = 4.353$

$$X(3) = \frac{3}{Z+R(3)} = \frac{3}{14.353} = 0.209$$

$Q_p(3) = X(3)\,R_p(3) = 0.209 \cdot 0.53733 = 0.1123$

$Q_1(3) = X(3)\,R_1(3) = 0.209 \cdot 2.6612 = 0.5562$

$Q_2(3) = X(3)\,R_2(3) = 0.209 \cdot 1.1547 = 0.2413$

$n=4$

$$R_p(4) = D_p(1+Q_p(3)) = 0.5 \cdot 1.1123 = 0.5561$$

$$R_1(4) = D_1\,(1+Q_1(3)) = 2 \cdot 1.5562 = 3.1124$$

$$R_2(4) = D_2\,(1+Q_2(3)) = 1.0 \cdot 1.2413 = 1.2413$$

$$R(4) = R_p(4)+R_1(4)+R_2(4) = \underline{\underline{4.91}}\ \text{sec}$$

$$X(4) = \frac{4}{Z+R(4)} = \frac{4}{14.91} = \underline{\underline{0.2683}}\ \text{sec}^{-1}$$

$$Q_p(4) = X(4)\,R_p(4) = 0.2683 \cdot 0.5561 = 0.1492$$

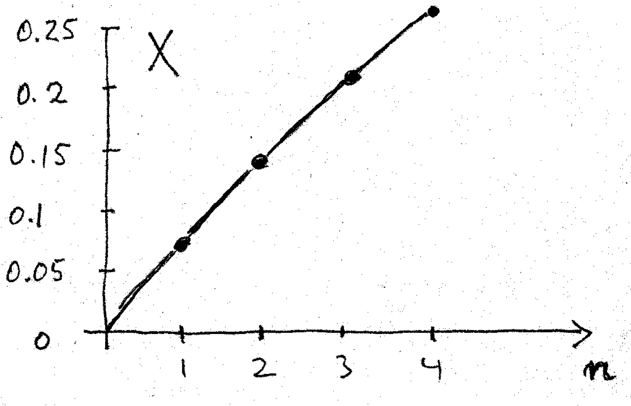
$$Q_1(4) = X(4)\,R_1(4) = 0.2683 \cdot 3.1124 = 0.8351$$

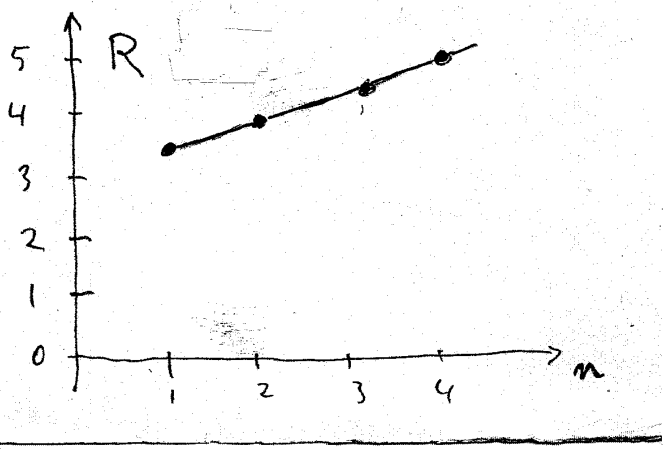
$$Q_2(4) = X(4)\,R_2(4) = 0.2683 \cdot 1.2413 = 0.333$$

$$U_p(4) = X(4)\,D_p = 0.2683 \cdot 0.5 = \underline{\underline{0.134}}$$

$$U_1(4) = X(4)\,D_1 = 0.2683 \cdot 2 = \underline{\underline{0.5366}}$$

$$U_2(4) = X(4)\,D_2 = 0.2683 \cdot 1 = \underline{\underline{0.2683}}$$


X
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
n


R
5
4
3
2
1
0
1
2
3
4
n