On the Design of Optimal Response Time in Computer Terminal Networks.

LCDR. An Young-Ki*

ABSTRACT

A terminal response time analysis for a general class of terminals-to-computer subsystem is presented in this paper. On the point of the front view, it should be considered for R.O.K. Military Defense to set up the communication network in order to facilitate for the currency of the information and the data communication system. The model used to study is based on the advanced data communications system in which terminals are connected to Terminal Control Units(TCU) that are in turn connected to local Front-End Processor(FEP). The line control procedures used to interface a TCU and an FEP may be half-duplex Binary Synchronous Communication(BSC), half-duplex Synchronous Data Link Control(SDLC), or full-duplex SLDC. This paper will contribute to facilitate the initial phase of system design and configuration for the Military Defense Communication Network System in future.

1. Introduction

System design of communication network is one of the great project confronted to us.

Design of the communication network has an advantage thus;

a. Exempted from storing of the overlapped data, that is, each computer center maintains independently and integratedly on their own usage. As we use the method of network systems by the terminals, we can save the cost and minimize the labor.

b. This system will contribute to increase the performance of the high speed data communication as to perform the logical computation by telecommunication.

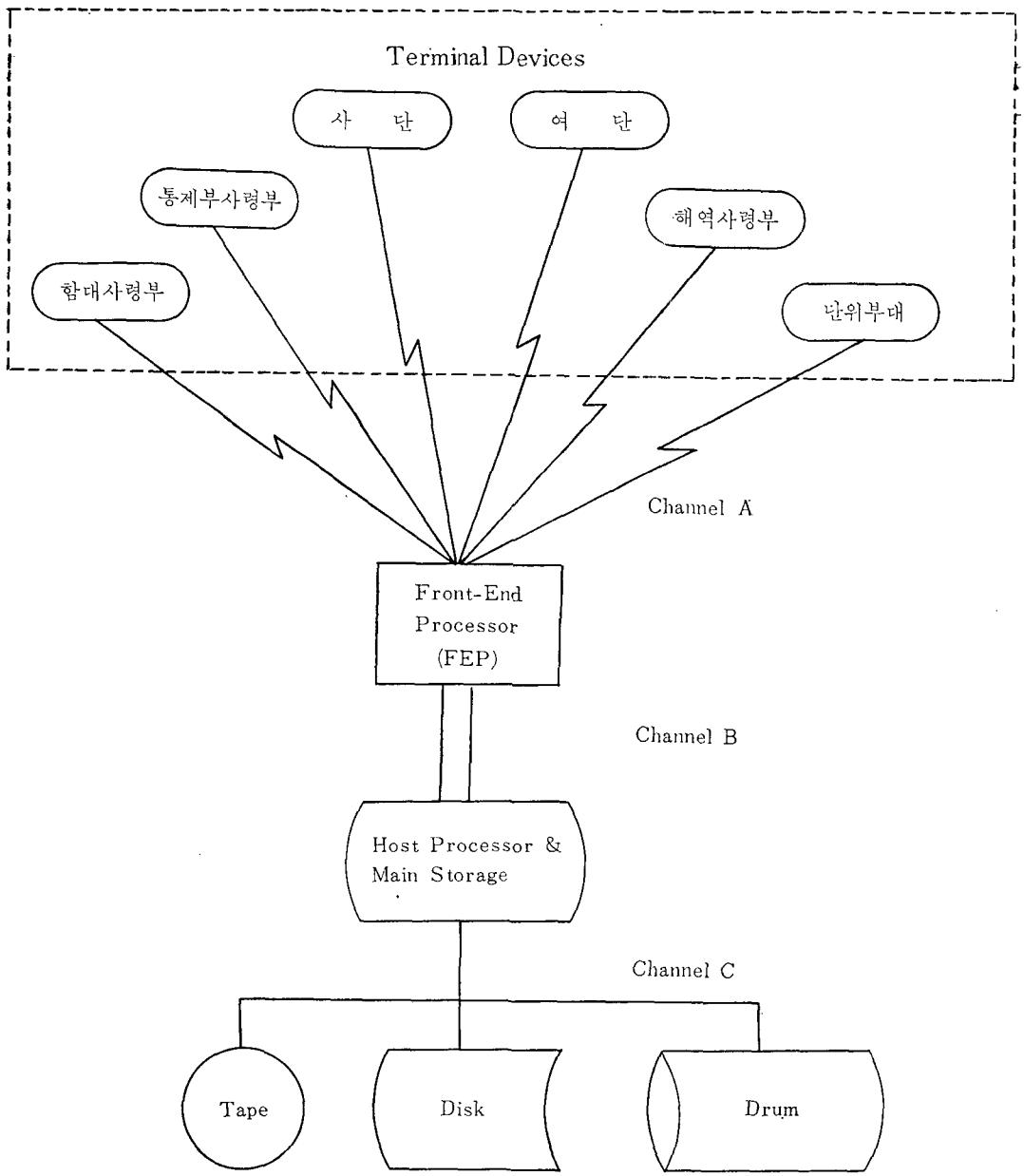
c. As not process at once, it doesn't require the lots of core size. Therefore, minicomputer performs as much as a large one. That is, it means minimize the cost of military budget. This paper is analyzed the little part in the entire system of the terminal network. A generic configuration of data communication system consists of many components such as Terminals, Terminal Controller Units(TCU), communications lines, remote as well as local Front-End Precessors(FEP), main storage, and auxiliary storage device(See Fig. A)

One of the important factor in design and evaluation of the network system is the analysis of the response time including turnaround time of the channel passage, which can be defined as the time interval from the operator's pressing the last key of the input to the terminal's typing or

* R.O.K Naval HQs.

displaying the first character of the response. Systems differ widely in their response time requirements, and the response time needed can, in turn, have a major effect on the design of the data transmission network and the data processing facilities. This paper presents the development of an analytical framework for analyzing response time requirements of data communications systems.

In case of setting up in one or two terminals in network system, it doesn't serious matter with the time to be consumed to process. But it is very important factor on the time when so many terminals is to set up to be needed for the data communication for the performance of the military



(Fig. A)

defense toward 1977 above.

The terminal response time as defined above is the totality of several time elements. At the time when the send key is depressed, the complete transaction has already been stored at a prespecified buffer area in the TCU, one for each terminal. Transactions stored at their terminal buffers cannot be transmitted to the host site until the particular TCU at which these transactions reside is polled by the local FEP in accordance with a given polling list. Not only kinds of the key board terminal, but also large terminal system (like as Univac terminal 9300 series in Naval HQs) shows a polling list on the printer with depressing the key-button by an operator the time spent by a transaction waiting for polling is the first time element to be calculated in obtaining the total terminal response time. This element depends on the system configuration and line procedures. The time required for transaction transmission along communications lines in relatively easy to calculate once we know the length of a transaction and the line speed.

When a transaction arrives at a FEP, certain delays may occur because this is where most communications functions are performed. After the whole transaction had entered the main processor, its processing time depends on the application programs, CPU processing speed, operating system, access methods, and the characteristics of the auxiliary storage devices such as disk files. A completely processed transaction will then wait at the FEP until the addressed line and TCU are ready to receive their responses. The length of this waiting time can generally be analyzed by an approximate queuing model.

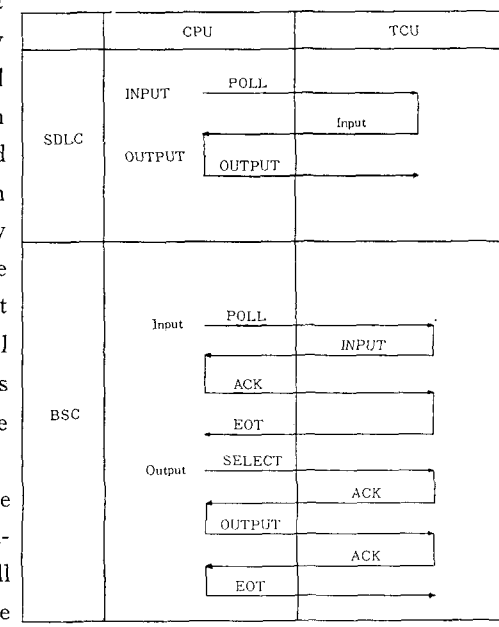
2. Cycle time for polling

Cycle analysis for operations.

Under the situation of the normal operatings serveral terminals as well as several TCUs may be prepared to transmit transactions at the same time from remote locations to the host site.

Only one can do so, and the others must wait their turns. To organize this, the line will normally be polled. For cases where terminals are controlled by the TCUs, as assumed in our model, transaction are sent to the controller at will and accumulated there so that only the TCUs need to be polled. In other cases, terminals are polled individually. Normally the local FEP or the host processor organizes the polling. In the main memory there is a polling list telling the programs the sequence in which to poll the TCUs or terminals. The polling list and its use therefore determine the priorities with which the remote devices are scanned.

There are several major time elements that constitute a polling cycle, and these include transaction transmission time, the time for either negative or positive poll and the associated acknowledgement. Except for the first item, all other elements depend to a great



(**Diagram A**)

extent on the line control procedures employed by the system. Two such procedures are considered in our model: the Binary Synchronous Communications(BSC) and the Synchronous Data Link Controls (SDLC). Some of the differences between these are that BSC can be used only for half-duplex transmission which SDLC can provide both half-duplex and the full-duplex modes, together with some built-in functions to reduce communication overhead as indicated in the follow diagram (Diagram A).

Here "POLL" is the polling message generated by the host site, "INPUT" is the transaction transmitted from the remote terminal to the host site, "OUTPUT" is the transaction transmitted from the host site to the remote terminal, "ACK" is the acknowledgement of the receipt of a transaction, "EOT" is the end of transmission, and "SELECT" is the selection of the proper remote device for receiving a transaction from the host site.

b. Transaction Transmission Time

Let K be a discrete random number variable denoting the number of transactions removed from a typical TCU each time this TCU is polled. The distribution function of K will be considered in the next section. We also let L_{in}, be the length (in number of characters) of each input transaction and S_L be the line speed (in number of characters per second). If there are Mc TCUs in the system, the total time required for the transaction transmission during a polling cycle is then

$$t_p^{(1)}=M_cKL_{in}/S_L \tag{1}$$

Although it is generally true that the sum of random variables may behave quite differently from each member of the sum, for analytic as well as practical reasons, we assume in Eq. (1) that all of the Mc terminal control units are identical in structure and all terminals generate similar traffic.

The extension to more rigourous cases is straightforward but would introduce many involved complications in computation.

c. Time element related to communications overhead

Let L_p, L_a, L_e be the lengths of polling, acknowledge and EOT messages, respectively, the time element related to a positive polling is then

$$t_p^{(2)}=\begin{cases} M_C(L_p/S_L+C_t), & \text{for SDLC} \\ M_C\{(L_p+L_a+L_e)/S_L+C_t\}, & \text{forBSC} \end{cases} \tag{2}$$

where C_t is a constant representing the time caused for by model establishment and other propagation delays.

In general the cycle time can be described in the form

$$t_p \doteqdot g(K)=t_p^{(1)}+t_p^{(2)}=\begin{cases} aK+b, & \text{if } K>0 \\ C_{np}, & \text{otherwise,} \end{cases} \tag{3}$$

where a=McLin/SI and $b=t_p^{(2)}$, and C_{np} is the time element associated with a negative poll. It is now necessary to obtain the distribution function of K.

d. Input process to the TCU

For the terminal subsystem under consideration, we assume that Mc TCUs are polled by a single FEP(there may be more than one FEP in large systems) and on the average M_t terminals are connected to a nearby TCU. Consider a particular terminal, one out of a group with a total of M_t terminals (the population source). In most interactive systems the terminal operator does not send

any inquiry before the response to the previous one has been received. Thus this terminal, after a time r_1, starts to transmit a transaction for the first time. After receiving the response, the terminal becomes idle for a time r_2 before making the second request for data transmission. In general, it stays for a time r_1 in the source before making the ith demand for the use of communications facilities.

Let the distribution function of the inter-arrival time of transactions at a TCU be

$$A(t)=\text{Prop}\ \ (T_i\leq t) \tag{4}$$

Instead of specifying the input process through the inter-arrival time distribution at the TCU from all the sources, which involves both the distribution and the size of the source, we specify the input process through the distribution of r_1 above, which is the inter-arrival time from one terminal.

Because the size of the source is M_C the distribution of $K(t)$, the number of arrivals up to time t is

$$\text{Prob}\ \{K(t)=k\}=\binom{M_t}{k}\ [A(t)]^k[1-A(t)]^{M_t-k} \tag{5}$$

which reduces to

$$\text{Prob}\{K(t)=k\}=\binom{M_t}{k}\ [1-e^{-\lambda t}]^k e^{-\lambda(M_t-k)t} \tag{6}$$

if the arrival process is exponentially distributed with parameter λ.

e. Number of transaction removed per poll in each TCU

The probability that k transactions have arrived at a TCU during a polling cycle t_p is

$$P_k=\int_0^\infty P(K(t_p)=k)f(t_p)\,dt_p \tag{7}$$

as t_p is a continuous random variable having probability density function $(p.d.f)\mathrm{f}(t_p)$.

On the other hand, t_p is also a function of the discrete random variable k as seen in *Eq.*(3).

It is thus proper to rewrite

$$P_k=\binom{M_t}{k}\ \sum_{l=0}^{M_t}\ [1-\exp^{\lambda g(l)}]^k[e^{-\lambda g(l)}]^{M_t-k}P_l,\quad k=0,1,2,\cdots\cdots M_t, \tag{8}$$
$$\text{and}\ \sum_{k=0}^{M_t}P_k=1,$$

where P_l has the same meaning as P_k with the subscript changed. We now have a set of simultaneous equations

$$P_k=\sum_{l=0}^{M_t}P_{lk}P_l \tag{9}$$
$$\sum_{l=0}^{M_t}P_l=1$$

with the coefficient P_{lk} given by

$$P_k=\binom{M_t}{k}\ [1-\exp^{\lambda g(l)}]^k[\exp^{-\lambda g(l)}]^{M_t-k} \tag{10}$$

It is noted that

$$\sum_{k=0}^{M_t}P_{lk}=1\ \text{for all}\ l,$$

so that these coefficients are transition probabilities.

The solution to *Eq.*(9) is given below by elementary manipulations

$$(P_0,\ P_1,\ ,P_{M_t}) = \frac{(1,\ P_1',\ P_2'\cdots,P_{M_t}')}{(1+P_1'+P_2'+\cdots+P_{M_t}')} \tag{11}$$

$$\text{where } (P_1',\ P_2'\cdots,P_{M_t}') = (P_{01},\ P_{02},\cdots,P_{0M_t})\ [I-(P_{lk}')]^{-1},$$

and where I is the identity matrix and (P_{lk}') is the $M_t \times M_t$ matrix formed by deleting the first row and first column of the original matrix (P_{lk}).

After having obtained the P_k, the nth moment of the polling cycle time can readily be expressed as

$$E(t_p^n) = \sum_{k=0}^{M_t} P_k[g(k)]^n \tag{12}$$

One can now fit $f(t_p)$ by a gamma distribution function

$$f(t_p) = \frac{\beta(\beta t_p)^{\alpha-1}e^{-\beta t_p}}{\Gamma(\alpha)}$$

with its Laplace transform

$$\Phi(s) = \int_0^\infty \exp^{-st_p}(t_p)\,dt_p = \left(\frac{\beta}{\beta+s}\right)^\alpha \tag{14}$$

where

$$\alpha = \frac{E^2(t_p)}{V_{ar}(t_p)} \quad \text{and} \tag{15}$$

$$\beta = \frac{E(t_p)}{V_{ar}(t_p)} \tag{16}$$

If $f(t_p)$ does not fit a gamma function, We can approximate $\Phi(s)$ by (4)

$$\sum_{r=0}^{\infty}(-1)^r\frac{s^r}{r!}E(t_p^r)$$

3. Waiting time for polling

The terminals in our model are assumed to be identical with respect to transaction generation intensity. The host processor (actually its local FEP) receives transactions and polls each TCU in a prescribed cyclic order (Polling list). Transactions that have been keyed in and are waiting at a given TCU are transmitted almost simultaneously after this TCU is polled. We fix our attention upon a givien simple terminal (out of the whole subsystem of M_t identical terminals) possessing a transaction generation intensity in number of transactions per unit of time and follow its history over a complete polling cycle. After the TCU is polled by the host processor, transactions waiting in the TCU are transmitted to the host site and the next TCU in sequence is served similarly. This particular TCU under consideration will be polled again after a random time t_p(polling cycle) and the host processor may find it either empty (negative polling) or with transactions waiting (positive polling).

Because the polling signal comes to any particular TCU every t_p seconds, where t_p assumes some known distribution, the polling process is indeed a renewal process. In particular, it is an ordinary renewal process because the t_p are independent identically distributed random variables. With the aid of some useful results available in the theory of the renewal process[4], we now evaluate the density function of the time that a transaction has to wait before being polled by the host site,

this situation is similar to a queuing process in which service is available only at service-intervals, which form a renewal process, A customer arriving at time t will have to wait a time t_w for the first service-instant. The limiting distribution of t_w can be expressed as

$$g(t_w)=\frac{1}{E(t_p)}\int_{t_w}^{\infty}f(t_p)dt_p \tag{17}$$

The moments of the limiting distribution of this waiting time are easily obtained from the Laplace transform.

Because

$$L[f(t_p);s]=\Phi(s)=\int_0^{\infty}\exp^{-st_p}f(t_p)dt_p=\sum_{j=0}^{\infty}(-1)^j\frac{s^j}{j!}E(t_p{}^j)$$

and

$$L\{g(t_w);s\}=L\left\{\int_x^{\infty}f(u)du;s\right\}=\frac{1-\Phi(s)}{s} \tag{18}$$

we have

$$L\{g(t_w);s\}=\sum_{j=0}^{\infty}\frac{(-s)^j}{j!}\frac{E(t_p{}^{j+1})}{(j+1)E(t_p)} \tag{19}$$

The jth moment of t_w about the origin, as it exits, is given by the coefficient of $(-s)^j/j!$ in the Taylor series expansion of its Laplace transform.

Therefore

$$E(t_w{}^j)=\frac{1}{j+1}\frac{E(t_p{}^{j+1})}{E(t_p)} \tag{20}$$

Three examples of different polling-cycle time distributions are considered here;

1) The polling-cycle time is exponentially distributed with $f(t_p)=\mu\exp^{-\mu t_p}$ where $\mu=1/E(t_p)$. The waiting-time distribution is easily shown to be

$$g(t_w)=\mu\int_{t_w}^{\infty}\mu\exp^{-\mu t_p}dt_p=\mu\exp^{-\mu t_w}$$

and thus

$$E(t_w)=\frac{1}{\mu} \tag{21}$$

$$V_{ar}(t_w)=\frac{1}{\mu^2} \tag{22}$$

2) The polling-cycle time is a constant. The density function is then the δ-function

$$f(t_p)=\delta\left(t_p-\frac{1}{\mu}\right)$$

Thus

$$g(t_w)=\begin{cases}\mu\left(0\leq t_w\leq\dfrac{1}{\mu}\right)\\ 0\left(\dfrac{1}{\mu}<t_w\right)\end{cases}$$

and

$$E(t_w)=\frac{1}{2\mu}=\frac{1}{2}\text{ cycle time} \tag{23}$$

$$V_{ar}(t_w)=\frac{1}{12\mu^2} \tag{24}$$

3) The polling-cycle time processes a gamma function density as given by E_q.(13).

Thus

$$g(t_w)=\frac{1}{\alpha}\sum_{k=1}^{\alpha}\frac{\beta(\beta t_w)^{\alpha-1}\exp^{-\beta t_w}}{\Gamma(\alpha)},\ \text{if } \alpha=\text{integer} \tag{25}$$

or

$$g(t_w)=\frac{1}{\alpha}\sum_{k=0}^{\infty}\frac{\beta^{\alpha-k}t_w^{\alpha-k-1}}{\Gamma(\alpha-k)}\exp^{-\beta t_w} \text{ if } \alpha\neq\text{integer} \tag{26}$$

with the meanand variance given by

$$E(t_w)=\frac{1}{2}\frac{\Gamma(\alpha+2)}{\alpha^2\mu\Gamma(\alpha)} \tag{27}$$

$$V_{ar}(t_w)=\frac{1}{\alpha^4\mu^2\Gamma(\alpha)}\left\{\frac{\alpha}{3}\Gamma(\alpha+3)-\frac{1}{4}\frac{\Gamma^2(\alpha+2)}{\Gamma(\alpha)}\right\} \tag{28}$$

4. Service time of the main storage

We cam consider the following assumption;

a. Ther is more than one program resident in the main memory, giving rise to contention among processing resources.

b. The CPU can be operated concurrently with the information transfer unit(ITU), which consists of the channel, the control unit; and the disk devices.

c. Both the queue in front of the CPU and the queue in front of the ITU are served under a FIFO (first-in, first-out) queuing discipline.

d. System overhead is negligible.

e. The number of programs line processed in the main memory, is a constant so that the system is a saturated mode.

f. The successive ITU service times are independently and identically distributed as a random variable W with arbitrary distribution

$$F_w(t)=\text{Prob}\{\text{W}\leq\text{t}\}$$

g. The successive CPU service times are independently and identically distributed as a random variable U with exponential distribution having rate parameter u,i.e.,

$$F_u(t)=\text{Prob}\{\text{U}\leq\text{t}\}=1-\text{e}^{-ut},\text{for } \text{t}\geq 0$$

h. A program requires a random number Mp of CPU services for completion and Mp has a geometric distribution with parameter q. The probability of termination after the j the CPU service is

$$\text{Prob } \{M_p=j\}=(1-q)^{j-1}q,\ j\geq 1$$

The effective Service time of the processor is then

$$E(S_H)=\frac{1}{E(D)}=\frac{\Pi_0+u(1-\Pi_0)\ E(W)}{q(1-\Pi o)u} \tag{E2}$$

Statistics of the ITU service time have been derived in (10) and Π_0 is calulated by

$$\Pi_0=\left[\sum_{i=1}^{N_{pr}-1}\gamma_i\right]^{-1} \tag{E3}$$

where $\gamma_0=1$, $\gamma_1=\gamma_0/G_0$,

$$\gamma_i=\frac{1}{G_0}[(1-G_1)\ \gamma_{i-1}-G_2\gamma_{i-2}-\cdots\cdots-G_{i-1}\gamma_1],$$

and

$$G_k=\int_0^{\infty}[F_k(t)-F_{k+1}(t)]dF_w(t), \tag{E4}$$

$k=0, 1, 2\cdots\cdots,\ N_{pr}-2,$ and

$$F_k(t)=1-\sum_{t=0}^{k-1}\frac{\exp^{-ut}(ut)_t}{l!} \tag{E5}$$

Suppose now that the ITU service tine has the Erlangian distribution, or

$$f_w(t)=\frac{\beta_1}{(\alpha_1-1)!}(\beta_1 t)^{\alpha_1-1}\exp^{-\beta_1 t} \tag{E6}$$

with $\alpha_1=\frac{E^2(W)}{V_{ar}(W)}$ (integer part);

$\beta_1=\frac{E(W)}{V_{ar}(W)}$ since $F_k(t)-F_{k+1}(t)=\frac{\exp^{-ut}(ut)^k}{k!}$

and $\int_0^\infty t^b e^{-at}dt=\frac{b!}{a^{b+1}}$

we can show that after some algebraic manipulations

$$G_k=\frac{\beta_1{}^{\alpha_1}u^k}{(u+\beta_1)^{k+\alpha_1}}\binom{k+\alpha_1-1}{l} \tag{E7}$$

If the above assumptions hold then, as shown in provious paragraph, the rate of departure from the host processor has the longterm expectation.

$$E(D)=\frac{q(1+\Pi_0)u}{\pi_0+u(1-\Pi_0)E(W)} \tag{E1}$$

5. Response time at the host processor

We can now calculate the response time at the host processor by treating the CPU and ITU together as a service facility. Because m many transactions enter and leave the host processor, dependence between various transactions seems small and the processor service time can be assumed to have exponential distribution with parameter E(D), given by Eq.(El). By applying an M/M/1 (Poisson arrival/Exponetial service time/simple serve) queuing model, we can get our estimate for the processor response time given by

$$E(T_H)=\frac{E(S_H)}{1-\rho_H},\ \text{and}$$

$$V_{br}(T_H)=\frac{E^2(S_H)}{1-\rho_H}$$

where $\rho_H=\Lambda_H E(S_H)$ and Λ_H is the total transaction arrival rate at the host processor.

6. Conclusion

As can be seen from the studies and researches, they are not entirely considered for the various factors affecting the terminal response time. By going through this paper, we can easily see that this time analysis is very complex and time-consuming. I believe, however, that this studies provided here furnish a outlined approach and framework for specific area.

It is the goal of this paper to give an "approximation" of the total terminal response time.

I can say the conclusive solution here;

(1) System analysis of terminal set-up for the military defense is absolutely necessary to examine this time analysis for minimize the budget of the Ministry of Defense.

(2) To set up the terminals in the military forces toward the military automation of the information and data communication is required of detailed and broad system analysis.

(3) Mathematical modeling approach to the actual performance evaluation.

I suggest that further effort be devoted to the validation and evaluation process in this field.

References

1. L. Kleinrock, "Analytic and Simulation Methods in Computer Network Design", *Proc. Spring Joint Computer Conference*, Spartan Books, New York, 1970, p.569.
2. B.V. Gredenk. avel I.M. Kovalenkl, "Introduction to Queuing Theory", *Israel Program for Scientific Translations*, Jerusalem, 1968.
3. Y.C. Chen and G.S. Shedler, "A cyclic Queue Network Model for Demand Paying Computer System", *Research Report, IBM* Thomas J. Vatson Research Center, Yorktown Heights, New York 10598, March 1970.
4. L. Takacs, "The Queues Attained by a Single Server", Oper. Res. 16, 639 (1968).
5. J.P Gray, "Line Control Procedures", *Proc. IEEE* 60, 1301 (1972)