

CUSTOMER QUEUING ARRANGEMENT

The present invention relates to a customer queuing arrangement and more particularly to such an arrangement for use in providing telecommunications
5 services.

Many organisations such as mail order companies, banks, repair companies, information providers and emergency service providers have many more telephone access lines than they have operators available to handle calls. This is a more efficient way of handling telephone access requirements since
10 operators are not idle while awaiting the next incoming call as would be the case if the number of operators and number of lines were equal. Further, at peak times, customers calling the service do not have to re-dial when lines are particularly busy.

It has become customary for an automated answering system to pick up
15 calls after a short period of ringing in order to re-assure the customer that they are connected to an appropriate service centre. Usually a message is transmitted such as "Please hold the line. You are held in a queue. One of our operators will answer your call as soon as possible." Music or other entertainment may be provided subsequent to the message or ring tone re-instated. After a period, often around
20 one minute, the message is repeated or another message is transmitted.

Some companies, notably in the United States of America, have introduced a "telephone radio show" in which customers joining a tele-sales queue are entertained by a disc jockey or comedian who periodically advises on the length of the queue and indicates the overall waiting time based on the duration of
25 the wait experienced by the most recently answered caller.

Such systems are helpful in avoiding lost calls to the destination but can lead to customer frustration if the wait is lengthy and the same message is repeated. This may result in lost business if the customer hangs up before being passed to an operator and subsequently calls again thus rejoining the queue.

30 The present invention seeks to overcome some of the above problems by providing a call queuing system in which the information provided is more specifically tailored to each customer's position in the queue.

According to one aspect of the present invention there is provided in a customer queuing arrangement a method of estimating an approximate time for which a customer in a queue may be required to wait for a server comprising the steps of determining the position of each customer in the queue, determining for
5 each of a plurality of active servers an average time to handle a customer, determining from the start time at which each active server last connected with a current customer the anticipated time at which such server will become available to obtain a series of times of anticipated availability of active servers and nominally allocating each customer in the queue to one of the series of times
10 whereby the approximate waiting time for each customer may be determined.

A preferred method of the invention comprises estimating an approximate time for which a designated customer in a queue of customers may be required to wait for a server comprising the steps of determining the number of customers (N) in the queue to be served before the designated customer, calculating the average
15 handling time for each of a plurality of servers to handle a customer connection to provide a mean service time (S), and multiplying the number of customers N by the mean service time S and dividing by the number of active servers (n) to obtain a wait time (W).

Preferably the calculated wait time W also includes an estimate of the
20 period at which the next server will become available. A further improvement of the calculated wait time may be obtained by comparing previously calculated wait times with actual wait times to derive a weighting factor adjustment for use in subsequent calculations of wait times.

Additionally, where customers may fall into differing categories a
25 respective mean service time may be determined for each of such categories and used in the calculation of wait time.

According to a second aspect of the invention there is provided a call queuing system comprising switch means having a plurality of input ports for connection to incoming telephone lines, a plurality of primary output ports for
30 connection to telephone handling servers, control means responsive to calls on the input ports to selectively effect connections through the switch means to one the output ports, the control means also being arranged to determine which, if any, of the output ports is not currently connected to an input port and to effect

connections sequentially to the output ports, the switch means further having at least one secondary output port to which calls arriving at the input ports are switched if none of the primary output ports are available, said secondary output port(s) being connected to an announcement system arranged to transmit a message to calling customers (unanswered customers), characterised in that the control means allocates each unanswered customer to the next position in a queue, calculates an approximate time for which the customer may be expected to wait prior to connection to a telephone handling server and causes voice announcement means to transmit the calculated time to the customer.

10 Preferably service data defines the average service time for each of a multiplicity of previously connected calls and server data defines the number of servers currently available such that the calculation of the approximate wait time takes into account the position of the connected customer in the queue, the number of servers available and the average service time.

15 Service data may be stored in respect of each of a plurality of different servers such that the calculated times may reflect variations in the speed of call handling by each operator.

The system may be arranged to store data defining the time of arrival of each call, the calculated wait time at the time of arrival and the actual time of connection of the call whereby a weighting factor based on the difference between the calculated and actual wait times may be applied to subsequently connected calls.

20 Where groups of input ports are allocated to different telephone numbers each of which may indicate a particular customer service requirement the calculation may include differing service times for each category of customer and/or for each telephone handling server.

A feature of the invention utilises historic records of caller activity to estimate potential waiting times taking into account expected stimuli or other predetermined factors.

30 Expected stimuli may include broadcast or published advertising, holiday or festival dates and the like.

Predetermined factors may include operator shift changes or breaks, an increase or decrease in the number of operators available and time-dependent

features such as optional extra discounts or other special offerings, and closure or opening of other server centres.

A call queuing arrangement in accordance with the invention using the method of the invention will now be described by way of example only with
5 reference to the accompanying drawings of which:

Figure 1 is a schematic diagram of the call queuing arrangement;

Figure 2 is a schematic diagram of the data tables used by the arrangement in Figure 1;

Figures 3a and b form a flow chart showing the handling of an incoming
10 call by the processors of Figure 1;

Figure 4 shows the calculate waiting time macro function of Figure 3b;

Figure 5 shows the calculate performance adjustment macro of Figure 3b;

Figure 6 shows the calculate handling time macro of Figure 3; and

Figure 7 shows a block schematic diagram of an adaptation of the PSTN
15 incorporating the customer queuing arrangement.

- Referring first to Figure 1 the arrangement essentially comprises a switch 1 having a multiplicity of input ports 2 for connection to incoming telephony traffic, for example for connection to the Public Switched Telephone Network (PSTN). A group of output ports 3 are arranged for connection to a number of
20 telephony servers for example telephone operators. A further port 4 is connected to a voice announcement system 5 and one or more ports 6 are connected to an entertainment facility, for example music on hold 7.

The connection through the switching system 1 is under the control of a store programmed controlled processor 8 which has access to a data store 9 the
25 contents of which will be described hereinafter.

The switching system 1 is arranged to selectively connect individual ports 2 to output ports 3, 4 and 6 under the control of the processor 8. Whenever a call is incoming to the system on the input ports 2, provided there is a server 3 free the system processor 8 will cause the connection of the respective port 2 to the
30 respective free port 3. First referring also to Figures 2 and 3, when an incoming call is received on any one of the incoming ports 2, the system processor at step 100 determines whether pointers (not shown) indicating the head of the queue and end of queue respectively in a queue table (table 1) of the data storage system 9,

are equal. If it is determined that the head and end of queue are equal the indication is that the queue for handling is currently empty. Thus the queuing system processor 8 interrogates the output ports 3 at step 110 to determine whether any of the servers which are currently active (as indicated in data table 2 of the data storage system 9) are free and if so obtains the server identity 115 of such a free server. The switching system is now arranged to connect the incoming call on the input port 2 to an appropriate port 3 at step 120. The call through the system is now under the control of the operator on the output port 3 and/or the caller on the input port 2 in the normal manner for handling of telephone calls through a telephony switch.

Having connected the caller to a server the system now identifies the category of incoming call (type) on the basis of the port 2 to which the call was first connected. Considering this identification further, it may be assumed that each of the ports 2 is allocated to one of a plurality of PSTN telephone numbers. Several ports may be associated with the same telephone number. However, the telephone number may indicate whether the call is to an order line for example or to a customer enquiry line. Alternatively, where a bureau is handling calls for several principles the telephone number may indicate the principle type of call. In a further development, the queuing system processor may use incoming line identification as provided by some system operators to facilitate the categorisation of calls.

Having determined the call type at step 125 the server stores the information to a particular server as shown in table 2. Thus referring to table 2 also, each server attached to the port 3 is allocated an entry in the table. Thus the destination port number for each server is represented in the left hand column and at step 130 the "real time" at which a call directed to the server is held in column 2 "LAST START". The type of call as identified at step 125 is also stored. Referring back to Figure 3 at step 135 the typical handling time of a call of a designated type is obtained from data table 4 and by use of the actual start time and the handling time an expected clear time can be entered into column 5 at step 140. In respect of any particular call step 145 "wait for end of call" now applies.

Returning now to the first step of handling an incoming call when at step 100 the head of queue is not equal to the end of queue thus indicating that one or

more callers is awaiting handling then the processor 8 performs a number of storage instructions in respect of the next position in queue.

Thus using the address in table 1 specified by the end of queue pointer prior to updating the pointer (step not shown) the system stores the location of the
5 caller at one of the locations 1-n in terms of the input port 2 on which the call has arrived. In other columns the "real time" at which the call arrived together with the call type as hereinbefore described with reference to step 125 are stored. A calculate wait time macro function 155 is now run in the processor 8 as hereinafter described and the estimated wait time and the time at which the
10 estimated wait was sent to the caller on input 2 is stored at step 160 respectively in the notice and estimate columns of the data table 1. As indicated at step 165 in respect of a particular call the processor now waits for a free server. Moving on to Figure 3b, if during the wait for a free server the processor 8 receives an interrupt this may be indicative of one of a number of events. One such event as
15 determined at step 175 is that a server on one of the ports 3 becomes active as free. If this is the case, at step 180 the processor 8 determines the head of queue from the head of queue pointer and at step 185 updates a further data table, table 3. The method of updating the data table is to remove from table 3 the oldest indication of estimated wait and actual wait and replace it with the estimated wait
20 from table 1 in respect of the caller now to be connected and the actual wait endured by that caller on the basis of the time of arrival of the call and the connection time. This data enables a performance adjustment to be carried out by a macro function (hereinafter described) step 190. The identified caller following updating the head of queue pointer (step not shown) is now connected at step 115
25 as hereinbefore described.

If at step 175 the system determines that there is no server free the interrupt may have arisen because a periodic timer, or real time indicator, shows that the period since notice was last given to the caller as indicated in table 1 has exceeded a predetermined period also. If this should occur the calculate wait time
30 macro function runs at step 195.

Other events which may cause an event interrupt may include additional servers coming on line or a change of server pattern or shift.

Once a call is in progress i.e. there is connection between an input port 2 and a server output 3 the processor 8 awaits an indication that the call has terminated and once the call is terminated proceeds at step 200 to determine from the start time as indicated in table 2 and the current time the actual time taken to
5 handle the specific call.

This information is stored in a further data table, table 4, in respect of the specific type of call as indicated in table 2 erasing the oldest data in respect of that particular call type. For the avoidance of doubt it is here noted that table 4 may be repeated for each server if variations in individual servers handling capacity
10 are required to be taken into account. It is also noted that where only one type of call or a single input type is identifiable then only a "single column" data table is required for each server or system.

Having stored the handling time the processor 8 may calculate the handling time at step 205. If the server remains active as indicated at step 210
15 then at step 215 a "server free" interrupt is provided to cause the next queuing call to move into step 170. Assuming that the server has decided not to handle further calls as determined at step 210 then table 2 is updated at step 215 by clearing the active indicator and until such time as the server again becomes active further action to connect input port 2 to output port 3 for the respective server will
20 not occur.

Considering the calculate wait time macro function as used in steps 115 and 195 of Figure 3 with reference to Figure 4 the current customer location in table 1 is marked (step 300) and using the head of queue pointer at step 305 the number of customers ahead of the current customer location in queue and their call
25 type is determined from table 1 and accumulated at step 310. For each type of call (referred to as TX) the average serving time (S_{av}) is obtained from data table 4 and multiplied by the respective number of customers ahead in the queue NX. To obtain the expected service time (S_x) for each of the types.

Using table 2 the number of active servers on the ports 3 is determined at
30 step 320 to provide a figure (n). Also from table 2 the expected clear time of the next server anticipated to clear can be obtained thereby obtaining a clear time (t_c) from which at step 330 an anticipated wait time W can be calculated as the sum of all of the expected service times S_x divided by the number of active servers (n)

minus the expected clear time period (tc). In order to make the calculations slightly more accurate a performance adjustment factor P may also be applied to the calculated wait time W.

Having calculated the wait time W the processor 8 causes the input port 2
5 on which the current customer is connected to be connected to the output port 4 to the voice announcement system 5 and at the same time passes the figure derived from W (that is the estimated wait time for the current caller) to be passed to the system 5. Thus the voice announcement system when connected to the caller may carry an announcement such as "you are currently held in a queue and
10 it is expected that an operator will answer your call in "W" minutes/seconds".

In dependence upon the point at which the calculate wait time macro was entered then the processor 8 returns either to await any further event interrupt or to storing the estimated wait time at step 160.

As has been previous noted as each call is cleared at step 145 the time
15 taken to handle the call is substituted for the oldest time to handle held in table 4 for the respective call type. This enables calculation by the macro 205 of the average handling time for the last "n" calls of specific type. Thus referring to Figure 6, the macro totals all the handling times for the last N calls of the particular type (if more than one) at step 350. By dividing this at step 355 by the number of
20 calls involved the average handling time of S_{av} can be stored in table 4 in respect of the particular call type. As has been previously indicated this figure is used in the calculation of wait time. Because the handling time is constantly being updated by replacing the oldest piece of data by the latest piece of data then variations of performance over a period of time can be taken into account by the
25 system so that the wait figure received by a caller on the input port 2 fairly accurately reflects current performance of the servers on the output port 3. As previously noted there may be a number of tables carrying server performance in respect of the call types. In general an average of each call type across all servers is calculated for use as S_{av} in the calculation of wait times. However, in respect
30 of expected clear time stored in table 2 it may applicable to have an S_{av} calculation in respect of each of the servers.

Turning now to Figure 5, to seek to make the best possible calculation of the wait time estimate (W) table 3 is used and each time a caller is connected to a

server on the output 3 at step 190 performance adjustment as shown in Figure 5 is carried out. Thus for each of the last "N" connections the estimated wait period is summed at step 400 and the actual wait period is calculated for the same N connections at step 405. The difference between the two calculations ($\Sigma E - \Sigma W$) is a measure of the actual performance of the estimating system and at step 410 a determination as to whether the system is over or under estimating the actual wait time is carried out. In dependence upon this performance factor P as used in the calculate wait time macro can be arrived at. Thus in one case at step 415 P is set at less than 1 that is to say the wait time to be broadcast requires to be shorter where in the other case at step 420 a factor greater than 1 is obtained.

Although a step to limit variations in the value of P has not been shown it will be appreciated that P may be bounded such that if $\Sigma E - \Sigma W$ at step 410 approaches zero no amendment of the current performance factor P is carried out.

The value of "f" shown in steps 415 and 420 may be system dependent and is determined by the operator on installation or subsequently.

It is noted that once a customer has received a call answer by virtue of connection to the voice announcement system 5 on clearance from the voice announcement the input port 2 may be connected to an output port 6 which provides some form of entertainment for example music on hold.

While as hereinbefore described the system is for a stand alone use with a multiplicity of exchange lines incoming to a limited number of operators, it will be appreciated that the system can be applied to a much larger network, for example the public switched network as shown in Figure 7. Thus referring now to Figure 7 the control system 28 may be incorporated as a platform on an intelligent network of the kind comprising a multiplicity of fully interconnected digital switching units here represented by digital main switching units 29 and 30. Each of the digital main switching units is connected to a number of local switches 31 by trunks 32 for providing service to lines to customer premises 33, 34, and 35. A network management system 40 controls the connection of the network. Interconnections between the MSUs may be as a result of data signalling using a specific data channel of time division multiplex linking through the digital network. This will usually use protocols such as that described in respect of C 7 signalling.

Now for example if an organisation has a number of lines 35 to its operators then calls from the PSTN customers for example those connected to lines 33 and 34 may be connected across the network only when one of the lines 35 becomes free and active. In this case, when a digital unit receives data from the local switch 31 indicating that a customer on one of the lines 33 has dialled a number for one of the lines 35 then the queuing control system platform 28 may be notified. If the control system 28 has received an indication by way of the DMSU 32 data channels that a free line in the server group is available then the network may be notified accordingly and the call switched though from the local switch 31 through the DMSUs to the local switch on which lines 35 exist and no further action, other than possibly to monitor for clearance as hereinbefore described for the purpose of calculating handling times will be required by the platform 28. However, if none of the lines 35 is currently free then the caller on line 33 may be locally connected to a voice announcement system 37 with the network control system providing periodic switching e.g. music on hold and voice announcement, with the expected wait being calculated and forwarded across the data side of the network. In this way the capacity of the public switched network is not used unless the call can be connected to the server group on one of the lines 35. Once one of the lines 35 becomes free the call queuing platform 28 may cause the system to set up the connection from the line 33 through the network to line 35 in a known manner. While a specific method of calculation of the estimated wait time has been described herein, another method of determining an expected wait time is to create a table of expected clear times in respect of each server which may be present (referring again to Figure 1) on the output ports 3. Thus table 2 carries an expected clear time based on the last start time and the type of call. Thus a table of types can be used to allocate each caller in the queue of table 1 on a nominal basis and by an iterative process determine the expected connect time for all subsequent callers in the queue.

The expected connect time may then be provided to the voice announcement system in the same manner as for "W".

While not specifically described herein, it will be noted that the system then responds specifically to variations in call dropout rate (that is callers who release from input ports prior to connection to a operator).

Further, the system may be pre-programmed by the operator with timings at which servers may join or leave the network server pool, for example and other factors such as seasonal variations may be taken into account.

In a further development, the system may be arranged to calculate the
5 expected waiting time W using more than one of the methods described previously so that utilising the performance adjustment P and comparing the estimated waiting time with the latest waiting time for each method of calculation, the most accurate current estimation method may be used.

Further factors which may be taken into account in the calculation of P or
10 in considering the weighting factor may be based upon historic records of calling patterns, weighting factors and performance accuracy for a corresponding period, for example at the same time of day on the same day of the preceding week or month or year.

The system may also be responsive to operator input of anticipated
15 advertising times, mail shots of press advertisements to determine the expected response and to modify performance adjustment accordingly.

Other factors, including anticipated shift changes, closure of other server centres and the like may also be taken into account.

The system may also be used to predict server requirements by utilising
20 historic records in combination with the waiting calculation methods so that by determining a simulated calling pattern the number of servers required to bound the estimated wait period may be estimated. Such information may be used by a service provider to vary the operation relationship pattern of the servers in use.

The system may be used to provide information to the operator for
25 example by setting a threshold for a maximum permissible wait time such that if in the first calculation of wait time as indicated at 155 of Figure 3a the anticipated wait time exceeds the predetermined threshold a supervisor position is notified.

Further, the calculate wait time macro of Figure 4 may be amended such that after the calculation of the wait (w) at step 330 the value of w is compared
30 with a number of thresholds and, in dependence upon the length of wait the message to be provided by the voice announcement is varied.

Thus the expected wait exceeds or is close to a near maximum threshold the announcement may advise the customer to disconnect and call at a later period

without insisting that the customer does so. Thus, connection may be to a particular voice announcement at step 335 of Figure 4 which provides a message such as "you are currently held in a queue and it is expected that our operator will take approximately 15 minutes to answer your call. You may wish to clear and

5 call back later when our lines will be less busy".

Where the wait is shorter for a wait of between one and two minutes then the announcement may be "please hold the line you are held in queue, one of our operators will answer your call within approximately w seconds".

**CLAIMS**

1. In a customer queuing arrangement a method of estimating an approximate time for which a customer in a queue may be required to wait for a
 5 server comprising the steps of determining the position of each customer in the queue, determining for each of a plurality of active servers an average time to handle a customer, determining from the start time at which each active server last connected with a current customer the anticipated time at which such server will become available to obtain a series of times of anticipated availability of active
 10 servers and nominally allocating each customer in the queue to one of the series of times whereby the approximate waiting time for each customer may be determined.

2. A method of estimating an approximate time for which a designated
 15 customer in a queue of customers may be required to wait for a server comprising the steps of determining the number of customers (N) in the queue to be served before the designated customer, calculating the average handling time for each of a plurality of servers to handle a customer connection to provide a mean service time (S), and multiplying the number of customers N by the mean service time S
 20 and dividing by the number of active servers (n) to obtain a wait time (W).

3. A method of estimating an approximate time as claimed in claim 2 further characterised by estimating the approximate time period [R] until one of the servers is next expected to be free such that

25
$$W = \frac{N \times S}{n} + R$$

4. A method of estimating an approximate time as claimed in claim 3 further characterised by determining for each of a number of previous customer connections the difference between an originally estimated waiting time (W1) and
 30 an actual waiting time (W2) to provide a weighting factor (P) relating to actual performance and adjusting the wait time such that

$$W = \left\{ \frac{N \times S}{n} + R \right\} \times P$$



5. A method of estimating an approximate time as claimed in claim 2, claim 3 or claim 4 further characterised by determining for each customer an expected customer category, determining the number of customers (N1,N2 etc.) in each category in the queue ahead of the designated customer, determining for each customer category a mean service time (S1,S2 etc.) and substituting $\{N1 \times S1\} + \{N2 \times S2\} + \dots$ for $\{NxS\}$ in the calculation of W.
6. A call queuing system comprising switch means having a plurality of input ports for connection to incoming telephone lines, a plurality of primary output ports for connection to telephone handling servers, control means responsive to calls on the input ports to selectively effect connections through the switch means to one of the output ports, the control means also being arranged to determine which, if any, of the output ports is not currently connected to an input port and to effect connections sequentially to the output ports, the switch means further having at least one secondary output port to which calls arriving at the input ports are switched if none of the primary output ports are available, said secondary output port(s) being connected to an announcement system arranged to transmit a message to calling customers (unanswered customers), characterised in that the control means allocates each unanswered customer to the next position in a queue, calculates an approximate time for which the customer may be expected to wait prior to connection to a telephone handling server and causes voice announcement means to transmit the calculated time to the customer.
7. A call queuing system as claimed in claim 6 further characterised in that the control means calculates an approximate time from service data defining the average service time for each of a multiplicity of previously connected calls, server data defining the number of servers currently in operation and the number of customers in the queue ahead of the customer currently connected to the voice announcement means.



8. A call queuing system as claimed in claim 7 further characterised in that the service data defines the average service time for each of a plurality of different servers and the server data defines the actual servers currently in operation.

5 9. A call queuing system as claimed in claim 7 or claim 8 further characterised in that each time a call is connected to the voice announcement means, the control means causes data defining the calculated waiting time to be stored together with the actual time of connection and upon connection of such call to a telephone handling server determines the difference (if any) between the
10 calculated waiting time and the actual waiting time such that a weighting factor may be applied (or adjusted) in subsequent call waiting time calculations.

10. A call queuing system as claimed in claim 7, claim 8 or claim 9 further characterised in that each input port is allocated to one of a plurality of telephone
15 numbers which determine the service category selected by calling customers, and service data is stored for each category such that in calculating the waiting time for a designated customer the control means determines the number of queuing customers ahead of the designated customer in each category together with the service data for their respective categories and the number of active telephone
20 handling servers to determine the announced waiting time.

11. A call queuing system as claimed in Claim 6 further characterised in that the control means determines from stored data defining the time at which each server was last connected to a telephone call and service data defining the average
25 length of time taken by each server to handle a connected call to produce a list of anticipated times at which each server will become free, the control means correlating the position in a queue of each unanswered customer with the list of anticipated times to determine the approximate waiting time.

30 12. A call queuing system as claimed in any one of claims 6 to 11 further characterised in that on occurrence of an event the control means causes each queuing customer to whom a previous announcement has been made to be



reconnected to the voice announcement system and re-calculates the waiting time to be announced.

13. A call queuing system as claimed in claim 12 further characterised in that
5 an event is the expiry of a pre-determined period of time since the previous connection of the customer to the voice announcement system.

14. A call queuing system as claimed in Claim 12 further characterised in that
an event is the connection of one or more preceding customers in the queue to a
10 telephone handling server.

15. A call queuing system as claimed in claim 12 further characterised in that
an event is a change in the number of active telephone handling servers.

15 16. A method of estimating call waiting times using the method of any one of
claims 1 to 6 characterised in that the queue of customers is simulated utilising
historic records of caller activity together with an estimate of the availability of
active servers and expected events.

20 17. A method as claimed in claim 16 further characterised by determining the
effect on caller activity of known stimuli and utilising the expected time occurrence
of such stimuli to determine future caller activity potential.

18. A method as claimed in claim 16 or claim 17 further characterised in that
25 the waiting time is limited by the operator whereby the required number of servers
to service an anticipated calling pattern for required limitation may be determined.