Robin Harris Foster and Stanny De Reyt

Re-inventing the Call Centre with Predictive and Adaptive Execution

Call centres have evolved from simple single-function centres to offer access, convenience, choice and courtesy to callers. Forecasting and staffing tools support planning, enterprise databases permit the business to craft specific caller treatments, and cross-trained agents using desktop applications can respond to a wider range of caller needs and business opportunities on a single call. One key element of the call centre, however, has changed only superficially—the question of 'What should each agent do next?'

The 'oldest waiting call' rule has answered that question for the last 20 years. Signs that this methodology is obsolete are seen in call centres where designs become more complex and results more difficult to achieve; where manual intervention moves agents from skill to skill chasing problems; where the most talented agents are overworked. This paper describes predictive and adaptive techniques† that answer the question 'What should an agent do next?' These techniques re-invent the call centre, creating a robust operation where performance is aligned with business intentions, without the manual, corrective intervention common in conventional centres.

What Should an Agent Do Next?

In any call centre, some routine must answer this question. First, what is the agent capable of doing? The agent's capabilities are referred to here as *skills*. Are the skills to be used in some hierarchical way, using some skills in preference over others? Are any of the calls waiting for the skills the agent holds queued at different priorities, where a call at a higher priority must be taken over a call at lower priority? Finally, how long has each call been waiting for service?

Robin Harris Foster:
Lucent Technologies, Room 1F-425
200 Laurel Avenue, Middletown,
NJ 07748, USA.
Tel: +1 732 957 5451
Email: robinhfoster@lucent.com
Stanny De Reyt:
Lucent Technologies Europe
Avenue Marcel Thirylaan 81
B-1200 Brussels, Belgium.
Tel: +32 2 7777 876
Email: sdereyt@lucent.com

This information can be combined to create three variations of the conventional oldest waiting call rule to assign a call to the available agent:

- the highest preference, highest priority, oldest waiting call is selected;
- the highest priority, oldest waiting call is selected; or
- the oldest waiting call is selected.

These rules alone may not achieve the results a centre seeks. Many call centres use call flow designs that change a call's queue priority or queue a call for one or more alternate skills in an effort to increase the chance of the call being served quickly.

The results of this type of control for a retail catalogue call centre in the United States are depicted in Figure 1. This diagram shows the daily average speed of answer (ASA) for sales and customer service calls. Although the centre provided an advantage to sales calls, customer service calls were too often handled sooner than sales calls.

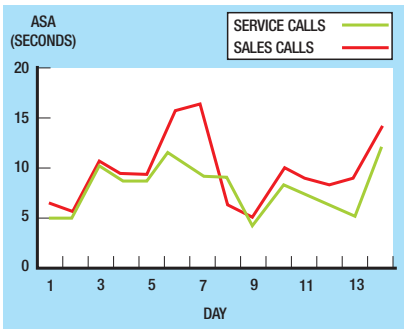


Figure 1—Call centre daily performance using conventional techniques

Re-inventing the Answer to What Should the Agent Do Next?

The use of the word *should* is deliberate in framing this question. Should implies some evaluation of alternatives or consequences in making a decision. What consequences are relevant in selecting a queued call?

New perspective on wait time

If two or more calls are waiting in the queue for the agent who has become available, the issue is one of call selection—deciding which call will be served. One element of consequence in this decision is the amount of time the calls *not* chosen will need to wait until other agents serve them.

A rare resource, if lost, presents a greater loss than a commonly found resource. If the available agent holds a skill that few agents handle, and

† The techniques described here were developed by Bell Laboratories, the research and development arm of Lucent Technologies. These techniques are embodied in software marketed by Lucent Technologies under the name CentreVu® Advocate and are the protected intellectual properties of Lucent Technologies.

another skill that many agents handle, the lost opportunity to use the rare skill must be considered. The understanding of how long a waiting call will continue to wait if passed up by the available agent is the foundation for building predictive and adaptive call centre operations.

A predictive evaluation of caller wait time

The amount of time a caller has waited and the amount of time a caller will remain waiting are both relevant in evaluating consequences. Bell Laboratories has developed a predictive algorithm, called *predicted wait time*: the sum of the caller's *current wait time* and the wait time until the next agent—not the one presently available—takes the call. The amount of time until the next agent is available is called *advance time*. Advance time is calculated for the calls the available agent might serve. The call with the greatest predicted wait time represents the neediest call among equally important calls.

Table 1 illustrates the evaluation of predicted wait time for queued calls. In this scenario an agent has become available and calls are shown queued for each of the agent's three skills—sales, customer service and inquiry. The calls arrived at various times and the current wait time for each call is known. The advance time calculation is the estimate of how long each call will continue to wait if not served by the available agent. The predicted wait time is the sum of the current wait and advance time for each call.

The call with the greatest predicted wait time in Table 1 is the customer service call. Notice that an oldest call waiting rule, which considers only the current wait time of the calls, would have selected the sales call. The sales call did indeed arrive three seconds before the customer service call, but a choice made according to predicted wait time avoids the longer wait the customer service caller would sustain.

Thinking about urgency

A consideration of only predicted wait time presumes that all calls are equally important. How long any call might wait for service, however, must be viewed in light of the urgency of the call. In a hospital emergency room some patients must be treated immediately. Others without critical injuries or illness will wait much longer. In a call centre, a 10-second wait may be much too long for one type of call but fine for another type.

Call centres have used both queue priorities and agent skill preferences to create attention for calls. Queue priorities and skill preference are both preemptive. The amount of attention a skill receives is dependent on the number of calls queued at higher priorities and on how many agents hold that skill at a preferred level.

A combination of queue priorities and skill preference levels seem practical in theory, yet when coupled with consideration of only the elapsed wait time (oldest call waiting) these designs yield fluctuating results. The centre depicted in Figure 1 (for example) used a straightforward design that provided advantage to sales calls.

A calibration of urgency

A calibrated and flexible method of evaluating concern for any caller's wait time would allow every type of call an agent handles to be considered while providing a meaningful distinction of calls in terms of relative urgency. A *service objective* or nominal gauge of good service can be defined for each skill and used as a comparison for the predicted wait time calculated for each waiting call.

Table 2 depicts the same call-selection scenario previously evaluated with the addition of service objectives or measures of urgency defined for each skill. A 15-second service objective is set for sales, a 20-second service objective for customer service, and a 25-second service objective for inquiry. These service objectives give distinct advantage, but not preemptive advantage, to sales over customer service and inquiry and to customer service over inquiry.

In this example, the call that will be served next is the one with the highest ratio of predicted wait time to service objective. The customer service call is selected through the effect of service objectives. A different business may value these calls differently and would choose different settings for the service objectives.

Creating alignment of service levels

The use of predicted wait time and service objectives in determining which call an agent should take aligns performance with business intent. Figure 2 shows the call centre in Figure 1 after adoption of these predictive and adaptive techniques. The sales calls have a consistently better ASA than customer service calls in alignment with the business' intentions. This centre used service objectives to bias call handling and obtained these characteristic 'heart beat' results. This alignment was accomplished, moreover, without real-time manual intervention on the part of call centre managers.

Identifying and Responding to Performance Problems

Call centres must perform well despite the random nature of call

Table 1: Using Predicted Wait Time in Call Selection

Skill	Current Wait Time	Advance Time	Predicted Wait Time (PWT)	Oldest Waiting Call	Call Selected
Sales	11 sec	9 sec	11 + 9 = 20 s	✓	
Customer Service	8 sec	24 sec	8 + 24 = 32 s		✓
Inquiry	3 sec	7 sec	3 + 7 = 10 s		

Table 2: Using Predicted Wait Time and Service Objectives in Call Selection

Skill	Service Objective (SO)	Current Wait	Advance Time	Predicted Wait Time (PWT)	Ratio PWT/SO	Call Selected
Sales	15 s	11 s	9 s	11 + 9 = 20 s	20/15 = 1·33	
Customer Service	20 s	8 s	24 s	8 + 24 = 32 s	32/20 = 1·6	✓
Inquiry	25 s	3 s	7 s	3 + 7 = 10 s	10/25 = 0·4	

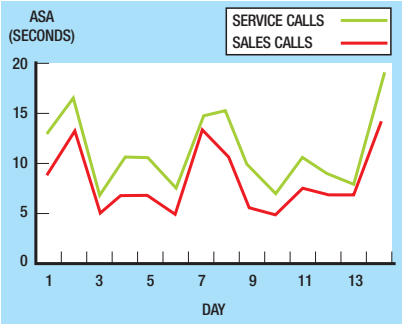


Figure 2—Call centre daily performance using predictive and adaptive techniques

arrival. Planning efforts, such as forecasting and scheduling, only prepare a centre for some baseline of operation. Deviations from the expected staffing, call volumes or mix or handle times for calls combine to create performance problems that conventional call-handling cannot overcome.

The manual intervention strategy

Call centre managers routinely try to adapt the real-time operation of the centre to maintain performance under unexpected conditions. These efforts follow a fundamental cycle. The manager must:

- watch the performance metrics of the centre,
- identify a problem condition,
- formulate a contingency solution,
- execute the solution,
- re-monitor the performance,
- determine when the problem has passed, and
- revert to the original method of operation.

This cycle takes precious time to execute. In addition, the metrics the manager observes are historical—a rolling ASA or the wait time of the queued calls. Caller wait times may already be too long before the manager can detect the problem. Often more than one problem must be solved simultaneously. Before a manager can return agents to their original skills the manager's attention might be drawn to the next problem.

A predictive and automated approach for adaptation

The technique developed by Bell Laboratories for managing performance in real-time allows potential problems to be identified and solutions applied before callers experience actual problems.

A method for predicting that a caller wait will become too long is to evaluate each caller's wait time as the call is placed into the queue and compare the prediction to one or more customer-defined *overload thresholds*. When a prediction exceeds the threshold, additional agents who could handle this type of call are activated or made eligible to serve that type of call. These *reserve agents* are part of the centre's contingency operation. When the prediction drops below the threshold, the pool of agents handling this type of call reverts to agents who typically handle the call under normal conditions.

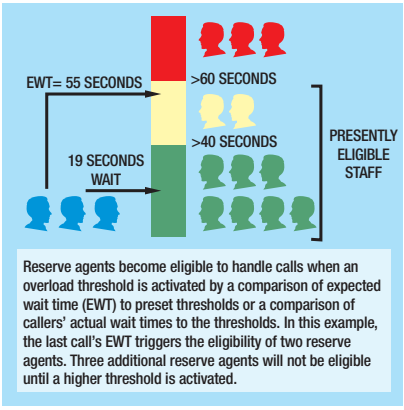
The prediction of a caller's wait time (used in contingency operations) is called *expected wait time*. The expected wait time calculation incorporates many dynamics of a call centre in its formulation. This predictive window into future call centre performance lets the centre's operation adapt to prevent calls from experiencing wait times above a desired threshold.

This concept is illustrated in Figure 3. A flexible system of thresholds allows a centre to align its automated contingency operations with its interests in eliminating long caller wait times. For example, an expected wait time calculation of 55 seconds for an arriving call is compared to the 40 second threshold set for the skill. Immediately, before this caller has experienced *any* of the predicted 55 seconds of wait, agents who hold this skill in reserve are eligible to handle this skill. The pace at which these calls are selected increases temporarily because the agent pool is larger. Multiple thresholds allow the reserve agents to become eligible in increasing numbers.

Perhaps only a few calls need to be served by this broader agent pool to reduce the potential for an unacceptably long wait time. When the expected wait time has dropped below the overload threshold, the contingency method of operations is halted and reserve agents are no longer eligible to handle this type of call. The thresholds can also be activated by the actual wait time of queued calls.

Reserve agent simulation

Table 3 shows the results of a very simple simulation using reserve agents. In this simulation, nine standard agents hold a single, standard skill where 110 calls arrive



Reserve agents become eligible to handle calls when an overload threshold is activated by a comparison of expected wait time (EWT) to preset thresholds or a comparison of callers' actual wait times to the thresholds. In this example, the last call's EWT triggers the eligibility of two reserve agents. Three additional reserve agents will not be eligible until a higher threshold is activated.

Figure 3—Activating reserve agents

per hour. The average handling time is four minutes. The performance results in terms of ASA, call completions and abandonments are noted. This simulation is compared to three scenarios using a 10th agent working first as a standard agent, then as a reserve agent with a threshold of 60 seconds, and finally as reserve agent with a threshold of 90 seconds. For simplicity in examining the contribution of the additional agent, the 10th agent is an exclusively reserve agent, meaning the agent has no other types of calls. In actual call centre implementations, however, reserve agents often handle other call centre calls.

Overall, any use of a 10th agent improves the throughput of the centre when measured as total completions per hour. However, the use of the 10th agent full time draws down the productivity of each agent by seven percentage points. When used as a reserve agent, however, the 10th agent's contribution has less impact on the standard agent's productivity.

A comparison of the additional calls completed to the amount of time the agent handled calls is one way to determine the beneficial leverage of the reserve agent. This is a measure of the leveraged contribution of the intermittent use of the reserve agent. In Table 3, the 10th agent, when working as a standard agent, is able to boost the completion rate by 1·8 calls/hour and is occupied 72% of the time. The reserve agent works with much greater leveraged contribution. Call completions rise by 1·4 per hour but with only 25% occupancy for the reserve agent with a threshold of 60 seconds and by 1·1 per hour with only 16% occupancy for the reserve agent. Considerable time remains for the reserve agent to perform other non-telephone duties.

Table 3: Reserve Agent Simulation Analysis (110 calls/hour, 4 minute average handling time)

Performance Measurement	Simulation Scenario			
	Baseline 9 Standard Agents	**10 Standard Agents**	**9 Standard Agents 1 Reserve Agent 60 sec threshold**	**9 Standard Agents 1 Reserve Agent 90 sec threshold**
Completions/Hour [A]	106·7	108·5	108·1	107·8
Completions /Hour over Baseline	Not applicable	1·8	1·4	1·1
Calls/Hour–Standard Agent [B]	11·9	10·9	11·6	11·7
Calls/Hour–Reserve Agent [C]	Not applicable	Not applicable	3·7	2·48
% Abandoned Calls	3·2%	1·5%	1·9%	2·2%
Average Speed of Answer	27·3 s	13·6 s	17·0 s	19·8 s
Maximum Delay (typical)	254 to 386 s	208 to 334 s	213 to 334 s	215 to 334 s
Maximum Queue Length (typical)	12 to 15 calls	10 to 12 calls	10 to 12 calls	11 to 12 calls
Occupancy–Standard Agents	79%	72%	77%	78%
Occupancy–Reserve Agent [D]	Not applicable	Not applicable	25%	16%
10th agent's leveraged contribution [E]	Not applicable	1·8 /0·72 = 2·5 calls/hour occupied	1·4/0·25 = 5·6 calls/hour occupied	1·1/0·16 = 6·9 calls/hour occupied

[A] Completions were greatest when the 10th agent held the skill as standard, but only marginally lower when the 10th agent worked less than 25% of the time as a reserve agent.
[B] Standard agents handle 1 less call/hour when the 10th agent is added as standard, but only 0·3 fewer calls per hour when the 10th agent is a reserve agent.
[C] The reserve agent handles more calls/hour at the lower, 60 second threshold but this work level does not appreciably lower the maximum delay over the contribution at the higher, 90 second threshold.
[D] The reserve agent's occupancy drops with the higher, 90 second threshold. In this simulation, the 10th agent is assumed to be at his/her desk, available for a call, but able to do other deskwork in the meantime.
[E] The leveraged contribution of the 10th agent is calculated by comparing the greater number of calls the centre can complete per hour divided by the occupancy of the 10th agent. This calculation magnifies the part time contribution of the low usage of the reserve agent.

Predictive and Adaptive Execution Each Time an Agent Becomes Available

If there are calls waiting for the agent to handle when the agent becomes available the various techniques described above are put into use sequentially to activate an appropriate contingency operation or to maintain normal operations.

If the agent has any reserve skills presently in an overload condition or any standard skills presently in an overload condition, business rules determine whether this call:

- is to be handled without consideration of any other calls waiting for standard skills;
- can be considered along with the other calls waiting for standard skills; and
- can be served only if no other calls are waiting for the agent.

If no reserve or standard skills are in an overload condition, a call is selected by examining the predicted wait times and service objectives for each waiting call:

- the predicted wait time is calculated for each call that the agent might handle;
- each waiting call's predicted wait time is compared to the service objective set for the skill; and
- the call with the greatest need, the highest ratio of predicted wait time to service objective, is selected and handled by the agent.

Distributing Work Fairly

Even in the busiest of call centres, there are many callers that do not have to wait for an agent. In some of these cases of agent surplus there are two or more qualified agents available. When a choice can be made, how the calls are distributed can create equitable workload for the agents, eliminating the 'hot seats' found in call centres.

When the *most idle agent* is selected from a group of available agents, the choice of agent is sensitive only to the amount of time since each agent completed the previous call. Selecting the agent who has had the least amount of call work—the *least occupied agent*—reduces the occupancy of busier agents and raises the occupancy of the less busy agents. The difference in these methods is illustrated in Table 4.

Often the agents who are the busiest are the agents with multiple skills. Their workload can be much higher than an agent who holds only one skill. The more skills an agent has, the more likely it is that at any time the agent becomes available there will be at least one call waiting for that agent to handle.

If a choice between two agents is possible and the least occupied agent

is selected, it is more likely the case that the agent chosen has fewer skills, perhaps only a single skill. Allowing a very busy, multi-skilled agent to remain idle has two important benefits. First, the multi-skilled agent has a longer period of rest, resulting in a lower occupancy and perhaps less stress over the course of the day. Second, during that extended period of rest, another call may arrive needing one of the skills that this available, multi-skilled agent holds. This coincidence allows a call that would have otherwise waited to be answered immediately.

Benefits of Predictive and Adaptive Call Centre Execution

The benefits of using predictive and adaptive techniques in call centres accrue to callers, to the profitability of the centre, to agent fairness and agent efficiencies, and to overall ease of management.

First, callers will experience fewer long wait times because predicted wait time allows the detection of long waiting consequences for individual callers. Service objectives bias call selection to provide appropriate advantage to different calls, bringing performance into alignment with business intentions. If sales calls should be answered sooner than service calls or top tier callers should be answered sooner than mid-tier, the service objectives allow this distinction to be made.

Sensitivity to the consequences of not choosing a call eliminates excessive delays, often causing maximum delays to drop. With fewer long wait times, abandonments often drop and more calls are completed. In a revenue-centric call centre each additional call handled raises revenue and reduces the caller's likelihood of calling a competitor. Each additional call completion eliminates a later retry and additional network costs, or perhaps a lost customer.

With fewer call abandonments, more calls are completed and agent productivity as measured in time spent on calls increases. The rules-based activation of reserve skills may reduce the number of back up calls agents take, potentially reducing call handle times. For example, agents may need to switch between transaction tools less often or the agents' increased focus on standard skills helps them proceed quicker or sell more on each call.

Table 4: Most Idle Agent Versus Least Occupied Agent as a Method of Agent Selection

Agent	Idle Time since last call	Work Time on ACD calls	Time since staffed-in	Occupancy (work time/ staffed time)	Most Idle Agent	Least Occupied Agent
Agent 1	30 s	50 min	60 min	83%		✓
Agent 2	35 s	90 min	100 min	90%	✓	

In actual call centre implementations, the automated mechanism for moving from routine to contingency operations has proven to eliminate real-time manual intervention. The time that managers had devoted to monitoring performance, planning and executing solutions and returning agents to standard operation can now be applied to more strategic issues of planning, coaching, and quality management.

Conclusion

The oldest call waiting rule, while practical in early single-function call centres, limits today's call centre in achieving performance levels aligned with the business' intentions. As the role of the call centre becomes more strategic to the concerns of the business, the call centre must incorporate the business' intentions more directly into its operations.

The predictive and adaptive techniques developed by Lucent Technologies and outlined in this paper reinvent the call centre at the core of its operations, with new considerations made each time a caller and agent are brought together. This re-invention permits the call centre to execute in alignment with its intentions in a consistent fashion and to eliminate corrective manual intervention.

Biographies

Robin Harris Foster
Lucent Technologies

Robin Harris Foster is the Manager of Research for Lucent Technologies, Call Centre Advocacy, and a former member of the technical staff of Bell Laboratories, the research and development arm of Lucent Technologies. Robin holds several patents in the field of call centres and multi-media call centres, including predictive and adaptive techniques offered as CentreVu® Advocate. She is active both in partnering with Bell Labs to advance the state-of-the-art in call centre technologies and in translating those advances into business-useful call centre applications for Lucent's customers worldwide.

Stanny De Reyt
Lucent Technologies

Stanny De Reyt has over 10 years of diversified experience in the telecommunications arena. He played a key role in starting the PABX business for Lucent Technologies (formerly AT&T) in Belgium and Luxembourg. He is known for his ability to apply innovative technology to meet customers' needs. Together with a large Belgian call centre customer he introduced to the Bell Labs scientists at Telecom Geneva in 1995 the concept of the *customer contact centre*. Two years later this call centre was able to provide to its customers the first Internet call centre applications outside the United States. His expertise spans a wide range of call centre technologies including automated call distribution, voice response, computer telephony integration, and outbound telemarketing systems. He has developed in-depth knowledge of the components necessary to implement and manage a successful call centre, especially in the insurance and financial industries and in the global system mobile operator business sector in Europe and the Middle East.