Developing a Management Model for a Complex Business System

C. Barnes[a], **D. Kulawiec**[b], **P. Marsh**[b]**. and R. Solomon**[c]

[a]Complex Systems Solutions, Canberra, A.C.T: [b] Australian Telco; [c] Taribon P/L.

Chris.barnes@brs.gov.au

Abstract: Due to the introduction of consumer service guarantee legislation, it became necessary for a telecommunications company to estimate their potential liability under the legislation and to devise a strategy to manage performance in order to reduce this exposure. Initial research indicated that the complexity of the business process prohibited detailed knowledge of the entire system by any one person. Using tools drawn from a number of quite disparate fields, including hydrology and manufacturing, techniques were devised to represent the end-to-end system performance in terms of the performances of a hierarchical combination of simple systems. Process representations of this kind were shown to be in a one-to-one relationship with a class of simulation models, which allowed accurate quantitative representation of sub-system performance at any level, including validation against observed data. Use of a single scaleable measure allowed comparisons of performance between geographically separate parts of the system regardless of size differences, and the identification of anomalies that might point to improved management strategies. Finally, use of the derived simulation models indicated possible strategies capable of improving performance by up to an order of magnitude for negligible cost.

Keywords: *process mapping; process management; complex systems; hierarchical process model; simulation*

1. INTRODUCTION

This paper describes the evolution of a methodology that was developed to solve a practical problem in the management of a complex business system. The problem arose because of the impending introduction by the telecommunications regulatory body, AUSTEL, of mandatory performance standards for consumer service, called Consumer Service Guarantees (CSG's). The standards were in the nature of time limits for particular types of service (telephone installations, rectification of faults) with initial standards being progressively tightened during the first year or so of operation. The standards also recognised that different standards should apply to different classes of customer in some cases, with the service to urban areas expected to be higher than to remote areas, for instance.

The initial question was: "Are the current service systems compliant?" Subsequently, a further series of questions of increasingly specific nature were posed, such as: "What are the major bottlenecks in the service system which cause delays?" and, "What standards are appropriate at sub-system level that will ensure whole-system performance?"

The approach taken by the consultants was to consider the problem as one of process management; in order to avail themselves of a considerable body of literature originating from the 1970's and '80's (for example, such seminal works as Deming (1986) and Senge (1990)) particularly in the USA and Japan. This approach, originally developed for the manufacturing sector, needed considerable modification in order to be applied to service industries- sometimes in quite subtle ways. For instance, the 'Motorola 6σ' approach to quality control (see, for example, Barney, 2002) is predicated on a normal distribution of component variations about the mean, whereas for human systems random variations are more typically exponentially distributed. The qualitatively longer 'tails' of the distributions require a somewhat different methodology, even though many of the principles are transferable.

The process improvement methodology chosen for this particular work was that of Integrated Process Management (IPM) (Slater, 1991). In the next section we briefly outline the pertinent features of this methodology. In the following section, the choice of an appropriate measure of system and sub-system performance is discussed, and the derivation of a system representation or process map is described. In section 4 it is shown that each form of process map gives rise to a simulation model, and example results are given. Finally, the conclusions will suggest lessons to be learned from this case study and further promising areas of application.

2. INTEGRATED PROCESS MANAGEMENT

Integrated Process Management (IPM) is a mature process improvement methodology that in essence consists of 6 well-defined steps including 2 feedback loops (Fig. 1).

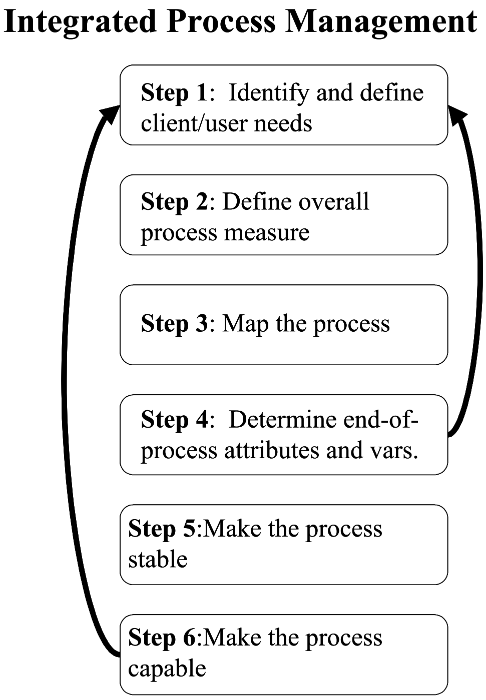


Fig. 1 Schematic representation of IPM methodology.

The first 4 steps are essentially the problem definition phase, which is iterated until the client is content that the problem being solved is the one required. The last two steps involve process improvement: firstly, by controlling process variation (analogous to the grouping of shots from a gun); and secondly, by ensuring that the process outcomes are centered appropriately on the required performance (analogous to centering the grouping on the bull).

The iterative nature of the methodology ensures that the initial configuration does not have a strong influence on the final result, and that dynamic systems can be tracked. It is a methodology to develop a structured procedure for designing, maintaining and improving processes. The framework itself does not guarantee success, but it does structure thinking about design that is integrated with process, management and corporate memory. Formal adoption of these concepts allows a fully structured and consistent approach to process characterization, monitoring and improvement. It is appropriate for application to complex systems because of its hierarchical approach to system management.

3. PROCESS MEASURE AND REPRESENTATION

3.1. Determination of Process Measure

The choice of process measure is the critical step in process improvement methodology, as it is this choice that largely drives the final outcome. Unless this choice is correct, it is quite possible to obtain unintuitive or even bizarre behavior as a result- this applies equally to human systems as well as deterministic or other systems with feedback. This marked dependence on the selected performance measure means that this is the point at which the greatest intellectual input and system understanding is required.

For the example considered here, after a lengthy process of consultation it became apparent that the nature of the problem suggested that a time-like measure was appropriate. The process measure chosen was that of the cycle-time, or end-to-end process time, distribution. The same measure could then be applied to both the system as a whole, and to individual sub-processes, which were identified.

3.2. Process mapping

Two separate processes were identified which were to be examined independently, although there was a small overlap in some aspects. The two processes were telephone installations and repairs, with different regulatory standards applied to each.

Initially, attempts were made to map the processes by utilizing the knowledge of the Process Owners and their subordinates. However, it gradually became apparent that although these people were very knowledgeable about particular parts of the system, there was no individual that had detailed knowledge of the whole. Furthermore, the perspectives of different people involved were not always compatible; nor were they necessarily appropriate to our purpose.

In order to resolve this problem, a number of computer systems that were used for monitoring and assigning tasks and workloads were examined, and databases that recorded information at appropriate levels were identified.

At the highest level, one such system recorded the time at which a particular job was created, and subsequently, the time at which it passed from one work area to another, until eventually the job exited the system. Far from being an almost linear system as was commonly believed, it turned out that the system was massively parallel, with a high rate of internal churn.

For example, in the Repairs process, jobs could potentially have any status of over 100 possible at any one time, but may return to that status a large number of times before completion. By listing the job status and time for a given job number consecutively, it was possible to determine the distribution of times for a particular transition from one status to another. It was also possible to determine the likelihood of a transition from one status to another. Ultimately, it became apparent that the NxN transition matrix (N>100, see above) was sparse, with strongly preferred pathways, with many forbidden transitions, or nearly so.

Using this information, it was possible to progressively coalesce different nodes, and simplify the process map. It was found that objective rules for such simplifications could be defined, and the process repeated. Finally the process could be represented by a series and parallel representation of relatively few compound nodes, with transitions characterized by the proportion of total jobs (which could be greater than 100%) that made the particular transition (see Fig. 2a). Finally, the transition could also be characterized by the proportion of the total time for all jobs that were associated with the particular transition (Fig. 2b).

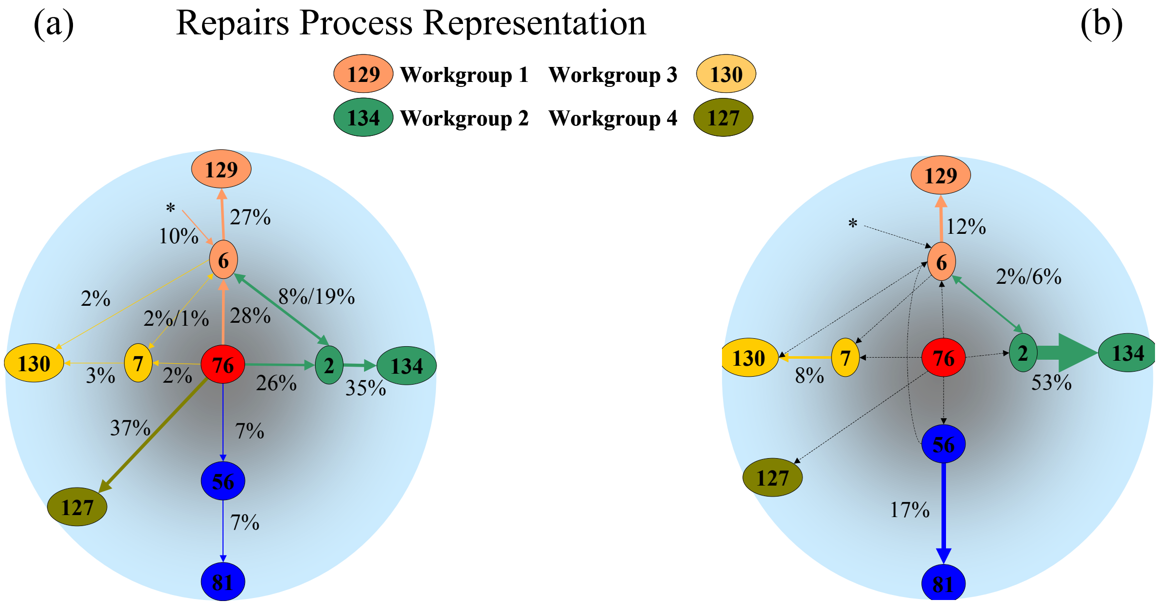


Fig. 2 High-level process map of Repairs process: (a) Representation by proportion of volume, (b) representation by proportion of total time (note that the sum of all transitions adds to 100%). The weight of the arrows is proportional to the volume or total time proportion, respectively, for the transition.

The net result was a hierarchical series/parallel representation of the process defined in terms of sub-processes identified (but not uniquely) *from the data.* Individual sub-processes could also be represented in this way, and so on to the limit of the available information. In some cases, alternative databases allowed a more detailed view still of critical parts of the system, and this was invaluable for zeroing in on process hot-spots.

The concept behind these diagrams is that jobs originate at node 76 in the centre of the globe (equivalent to 100% volume in the figure), except for 10% that are created by Workgroup 1 at node 6. Jobs "diffuse" to the edge of the globe, and exit through one of 5 possible routes. Although the pathway to node 134 accounts for only 35% of all jobs, the compound transition from node 2 to 134 accounts for 53% of all time. This suggests that this part of the process is time-critical, and requires further investigation. This procedure can then be repeated by expanding the transition representation

4. PROCESS SIMULATION

4.1. Model construction and validation

In order to explore possible management strategies, a hierarchy of dynamic simulation models were constructed from the process maps; one model for each representation. Each transition between nodes is associated with a transition time distribution function, determined from the data and regarded as the output from the next lower sub-system represented by the transition. By combining these distributions in the manner suggested by the process map, the observed end-to-end distribution should be re-created from the component distributions, thus validating the model at this level from data obtained at the next lowest level.

For pathways that are not time-critical, such as between nodes 76 and 127 in Fig. 4, there is no need to explore this pathway further. Otherwise, by drilling down, critical pathways can be isolated and examined in detail

The result is a hierarchical model which focuses only on areas of interest, and is consequently efficient. Model validation occurs at each stage, starting at the highest level, so that errors due to ignored transitions or approximate representations of distributions do not propagate.

4.2. Performance comparisons

Any distribution function can itself be represented as a series and parallel combination of exponential distributions or linear "reservoirs" (analogous to representing a unit hydrograph in the same way; e.g. Jakeman et al., 1990) with an arbitrarily high degree of accuracy. This implies that it is possible to use the above approach to objectively compare the performance of the system at different times (e.g. to measure the effect of management initiatives), or different subsystems simultaneously. This is a consequence of our choice of performance measure as the distribution of transition times; a measure which is (largely) independent of the size of the system, and expected to be fairly stable with time under similar conditions.

Fig. 3 shows the results of comparisons between mean cycle-times for different sub-systems, and the contributions of major component sub-systems for the somewhat more detailed level 1 system representation. The figure shows that all regions have a similar number of transitions, averaging nearly 3 per job, but that the time per transition varies by as much as a factor of 2. The contribution of individual transitions can also be distinguished, and it is clear that variations between regions arise principally from variations in just a few transitions, rather than all transitions contributing proportionally. This suggests either differences in conditions between regions, or differences in management practice, both of which are potentially of interest to managers.

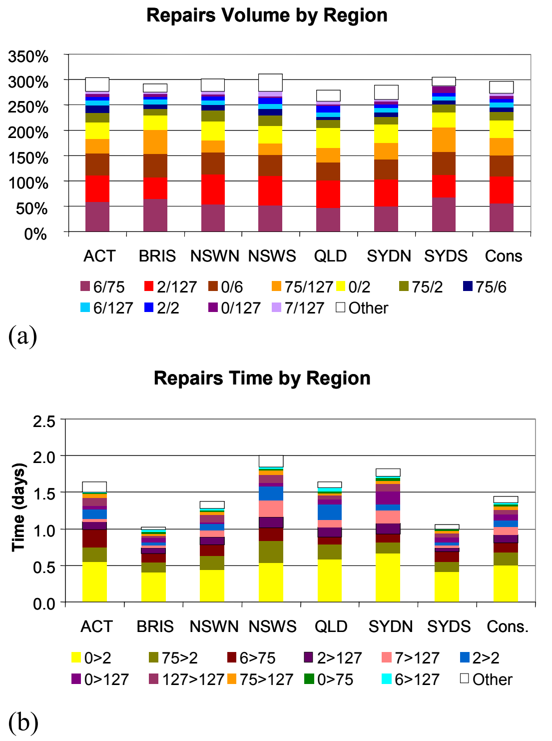


Fig 3. Level 1 transitions. (a) Regional breakup of mean number of transitions per job; (b) Regional breakup of mean time per job.

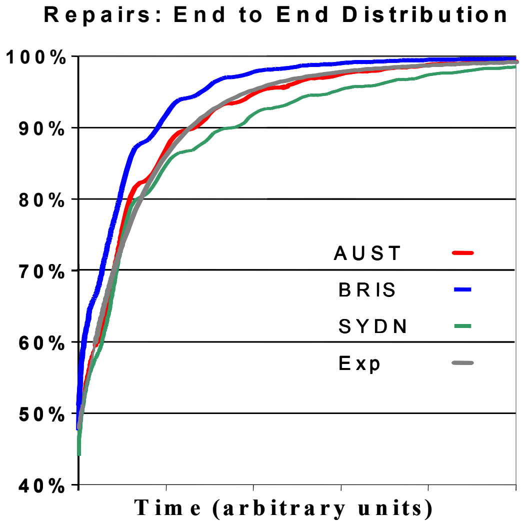
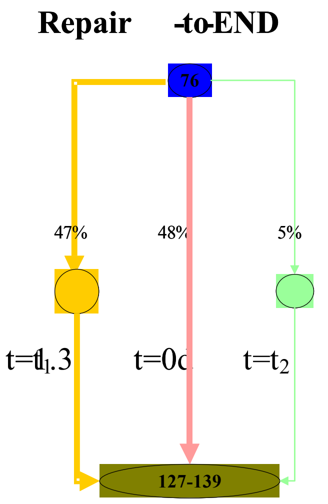


Figure 4. Cumulative distribution functions for 7 regions (only extreme distributions shown). Also shown is the consolidated distribution function for all areas together (red curve), and the exponential function which gives the best fit to the consolidated curve. The diurnal ripple is especially marked at short times.

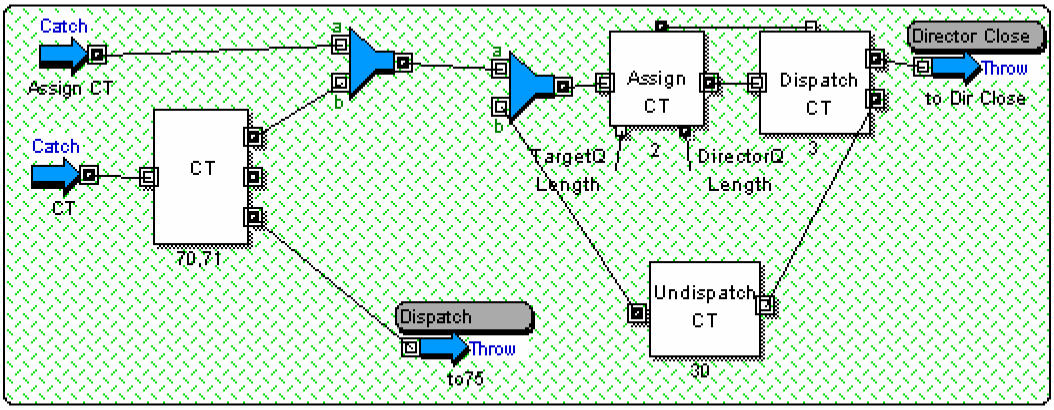
Figure 4 shows a comparison of cumulative cycle-time distribution functions for the end-to-end Repair processes between regions. Because the distributions are approximately of modified exponential form, and are highly skewed, the mean repair time is around the 80th percentile, where there is a factor of about 2 between the worst and best repair times.

Although this methodology provides a means of comparison between the performance of different systems, or the same system at different times, the comparison is not fully objective, depending on the calculation of a number of statistics which are not necessarily meaningful in terms of the systems themselves. However, having observed that any distribution function can be written as a linear combination of series and parallel queues or reservoirs, it is possible to objectively decompose the observed cycle-time distribution. For this class of structures, the decomposition is unique. Figure 5 shows the results of a decomposition of the observed Repair process cycle-time distribution (fig. 4) into an equivalent parallel sum of 3 simple queues, giving an accurate representation of the original distribution. This type of analysis shows that the process behaves as if there were three main pathways by which jobs could progress, each with its own characteristic time constant, and proportion of the total.



$$F(t)=A+B[1-\exp(-t/t_1)]+C[1-\exp(-t/t_2)]$$

Figure 5: Analytic decomposition of Repair process cycle-time. Approximately half of all jobs appear to behave as if they are in an exponential queue (random mixing) of timescale t_1 days, and another half are dealt with instantaneously. A small proportion (5%) behaves as if they are in a random queue of characteristic t_2 days, which then dominates the long-time behavior of the combined system.



:

Figure 6 Part of Extend© model for simulation of field technician assignment and queue.

4.3. Applied Simulations

Having identified the composite transition from node 2 to 134 as time-critical, simulation models were constructed to explore different management strategies that may reduce the problem associated with the transition. At this stage it was necessary to introduce data of higher process definition than the first from a separate database. After some simplification, a model of the structure of Fig. 6 was found to represent the observed data satisfactorily (see Fig. 7). Analysis of observed performance indicated that current queue management, with a complicated set of priorities and other factors, meant that on average the queue acted, not like a FIFO (first-in-first-out) queue, but as a random or Poisson queue, where the probability of exiting was independent of the waiting time in the queue. Through introducing a more aggressive queuing protocol, it was possible to show that average queue lengths could be reduced by an order of magnitude for negligible decrease in worker utilization (increase in cost), by adapting this different queuing strategy and giving all jobs equal priority.

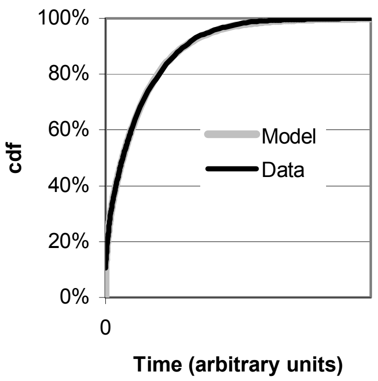


Fig. 7 Correspondence between measured end-to-end time distribution for the composite transition between node 2 and node 134, and simulation output from the model of Fig. 6 (5 parameter fit)

5. CONCLUSIONS

The original questions: "What are the major bottlenecks in the service system which cause delays?" and, "What standards are appropriate at sub-system level that will ensure whole-system performance?" can now be answered using the methodology developed above. Bottlenecks are defined with any desired resolution from the series of hierarchical time and volume process map pairs obtained directly from system data. The problem of ensuring an overall performance standard by managing the performance of component subsystems has been converted to a cascade of simple problems, each of which can be addressed in turn in a "top down" approach.

The choice of performance measure as the cumulative distribution of process cycle times allows a number of simple but powerful analytical techniques to be employed. The decomposition of the distribution function into a series/parallel sum of simple queue/servers allows the exploration of various management options. This method of decomposition suggests two main ways of process improvement: time-scales may be reduced through more efficient sub-processes; or the proportional contributions of different pathways can be altered by system re-design so that dependence on the slower pathways is reduced.

For each level of process mapping, a corresponding simulation model can be determined from the map, with validation coming directly from the data used to generate the map itself. Concentrating detail only where the overall behavior is sensitive to it results in an efficient compound model that closely mimics the observed behavior of the whole system, or subsystems at any level.

Although this methodology has been developed for a particular application to a telecommunications network, many of the techniques appear to be generic, and applicable to a wide range of complex service systems, possibly using performance measures other than time distributions as appropriate.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge additional members of the original team which carried out the work described here: John Clarke, Steve McManus, and the Process Improvement Team in the Telco; and Barry Newell of the Australian National University.

Extend© is generic simulation software from Imagine That Inc., CA, USA.

7. REFERENCES

Barney, M., 2002, *Motorola's Second Generation*, Motorola Inc., http://mu.motorola.com./pdfs/Mot_Six_Sigma.pdf, © American Society for Quality, 2002.

Deming, W.E., 1986, *Out of the Crisis*, Centre for Advanced Engineering Study, MIT

Jakeman, A.J, Littlewood, I.G. and Whitehead, P.G. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *J. Hydrol.* 117: 275-300.

Senge, P.E., 1990, *The Fifth Discipline, ISBN 0-385-26095-4*

Slater, R., 1991*, Integrated Process Management: A Quality Model*, McGraw Hill, NY.