2005

Management of Extremes in the Configuration of Interoffice Telephone Switch & Priority Systems

C. Ariel Pinto
Old Dominion University

Original Publication Citation

Pinto, C. A. (2005). Management of extremes in the configuration of interoffice telephone switch & priority systems. *26th Annual National Conference of the American Society for Engineering Management 2005 - Organizational Transformation: Opportunities and Challenges, ASEM 2005;, Virginia Beach, VA; United States; 26-29 October 2005* (pp. 166-170). American Society for Engineering Management.

MANAGEMENT OF EXTREMES IN THE CONFIGURATION OF INTER-OFFICE TELEPHONE SWITCH & PRIORITY SYSTEMS

C. Ariel Pinto, Ph.D., Old Dominion University

Abstract

This paper describes how to enable diverse enterprise customers for voice-data switch to achieve in configuration a balance among users, features, and perceived reliability subject to extremes of traffic. The analysis entailed the simulation of the voice-data switch with embedded priority system, generation of latency times for various configurations and transaction traffic rates, and the development of a framework and theoretical propositions for configuration of super-saturated systems. It was shown that the concept of tolerance levels defined in the risk of extreme events can be applied for embedded priority systems and was the basis for the application of the zone-configuration evaluation diagram.

Introduction

An inter-office telephone switch (or simply *switch*) can be defined as a device that connects inlets to outlets by concentrating call traffic where the inlet can be the calling party and the outlet can be the called party. A voice-data switch plays a major role for enterprise customers using multiple telephone lines and is the heart of a call center. Important qualities of a switch are functions, features, cost, and reliability. Switches offer programmable functions such as call-transfers, intercom, and data-capture meant to add conveniences and support to organizational functions.

This paper presents a management approach to configuring a switch for diverse types of users. This approach uses simulation and an evaluator diagram to provide managers with a way to technically evaluate a switch and to present the results to end users. This approach can be readily applied to other types of systems in IT, manufacturing, and service systems. Furthermore, two propositions regarding the latency times for super-saturated systems are presented. This provides managers insights to dealing with extreme loads during system operations.

Configuration addresses how many and what types of users are served by a single switch. There is a tradeoff in configuration between the prevalence of high-bandwidth devices such as large-screen telephones and the user-perceived failures of the local-switch system. The acceptable rate of failures varies among enterprises, and any single enterprise is concerned with various traffic extremes. At the heart of the switch is the priority system used to determine which messages, tasks, and processes are serviced by the switch resources such as processors and storage spaces.

Priority systems (PS) are composed classes of transactions competing for the service of one or more servers. As transactions arrive in the system, they are labeled according to some attributes such as time of arrival, origin, service requirements, or according to state of the system (e.g., number of waiting transactions, and server utilization). The labels are then used to distinguish the transactions into classes. The classes of transactions are assigned priority levels that represent relative importance and provide the basis for choosing the class from which the next transaction to be served will be coming. Exhibit 1 is a schematic diagram of a PS showing (from left to right) the stream of arriving transactions, the grouping of transactions into classes, and a server with a transaction currently being served (adapted from Pinto and Lambert (2002)).

Exhibit 1. Schematic Model of PS Showing Arriving Transactions Grouped Into 3 Classes With a Single Server Processing a Transaction From Class 3.

As such, one challenge for the switch operator is how to manage the extremes in operating such switches using the knowledge of the PS embedded within a switch.

Simulation Model of a Switch

We used a discrete-event simulation of switch reliability, identification of traffic scenarios, and multi-objective evaluation of alternative configuration. An objective of the simulation model is to characterize the reliability of a call center through off-hook latency times for different switch configurations and traffic rates. Specifically, the configuration includes functions programmed in the switch and hardware the switch is servicing (e.g. speed of the main processor in the telephone switch and number of small- and large-screen). Included in the sets of configuration is a particular model of PS configuration which determines the relative services of classes of transactions, called DPS1 (Pinto and Lambert 2002). Latency times longer than a specified threshold are considered as failure of the system. The relevant latency under consideration is from the occurrence of the off-hook event to the user perception of a dial tone. The latency involves the creation and routing of inbound messages and the generation and routing of outbound messages. Important parameters of the simulation model are:

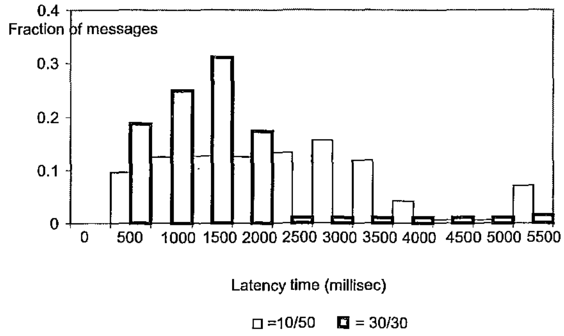
- Latency times of off-hook events
- DPS1 parameters
- Off-hook event rate
- Length of time between each occurrence of the small-screen (large-screen) telephone events.
- Length of sustained traffic rate is the time duration of a given traffic rate, to represent a burst of demand.

More details of the simulation are discussed in Pinto and Lambert (2003).

Result of the Simulation Analysis

Latency times. The simulation is used to generate the latency time for each of the simulated off-hook events using a set of simulation parameters described in the previous section. Such records can be depicted in histograms-- for a particular configuration of the switch f, a histogram of the latency times for a specific traffic rate-T can be obtained. Consider the decision regarding the number of large-screen telephones in a set of 60 telephones to be configured in a switch. Shown in Exhibit 2 are the histograms of latency times for two configurations, one for a switch with 10 large-screen phones and 50 small-screen phones and another for a switch with 30 large-screen phones and 30 small-screen phones using similar traffic rate of 500 calls per hour.

Exhibit 2. Latency Times.



Charts similar to Exhibit 2 are generated for all configurations that are of interest to dealers and users. However, the acceptable value of latency time depends on the telephone functions relevant to switch users (e.g., dial tone, LCD refresh). Each function can have an associated latency-time threshold; for example, a telephone function n with a threshold C_n, a latency-time x_n is unacceptable if $x_n > C_n$. Furthermore, let $R^j_n(C_n)$ be the fraction of latency times beyond threshold C_n for configuration j:

$$R^f_n = \frac{\text{Number of } x_n \text{ such that } x_n \geq C_n}{m}$$

Shown in Exhibit 3 are plots of the fraction of messages Rjn(Cn) in the vertical axis and configuration f, different numbers of large-screen telephones in a set of 60 in the horizontal axis for two thresholds C1 = 0.5 sec. for dial tone and C2 = 0.3 sec. for LCD refresh. A single traffic rate T = 500 calls per hour is used.

Exhibit 3. Fraction of Messages Exceeding Latency-Time Thresholds C1 =0.5 and C2 = 0.3.

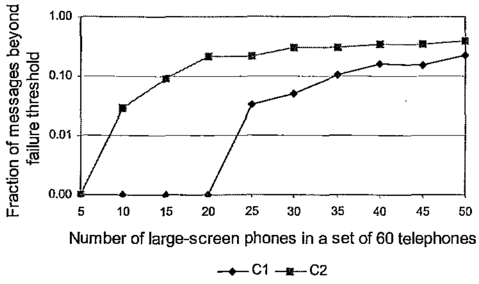


Exhibit 3 shows the fractions of messages with latencies exceeding the threshold when there are five large screen phones ($R^5_1 = R^5_2 = 0$), which means that a switch with five large-screen phones produces short latency times and no latency times are beyond the two thresholds. On the other hand, $R^{10}_1 = 0.28$ and $R^{10}_2 = 0.38$. From the perspective of the user, a switch with fifty large-screen phones may offer more capabilities than a switch with five large-screen phones, but latencies on specific telephone functions Class1 (dial tone) and Class 2 (LCD refresh) are also more frequent. A tradeoff in configuration between functions and performance is thus evident.

The process described to create Exhibit 2 and Exhibit 3 is repeated for different traffic rates T that is representative of switches used for up to 50 telephone units (100, 1000, 5000, and 10000 calls per hour). The choice of T was based on expert opinions of switch dealers and enterprise customers from different fields of applications. The charts are aggregated into a single table that shows the R^f_n, f, T, and C_n. One of the three variables f, T, and C_n is held constant.

Tolerance levels. An important consideration for a switch user in choosing the proper configuration is the intended application. For example, a business office may tolerate an unacceptable latency time ($x_n > C_n$) up to 1 of every 100 messages ($R^j_n(C_n) = 0.01$) while an emergency care application can tolerate only up to 1 out of 1000 messages ($R^f_n(C_n) = 0.001$). A lower bound on the fraction of messages $R^f_n(C_n)$ can also be imposed to prevent the underutilization of the switch. Thus, different fractions of messages $R^f_n(C_n)$ can differentiate the various fields of applications and leads to the definition of a variable that distinguishes the different bounds of fraction of messages $R^f_n(C_n)$ for various fields of applications. Tolerance levels are introduced by Pinto and Lambert (2002), and utilized for graphical presentation of configuration tradeoff in developing the (ZCED) Zone Configuration Evaluator Diagram (Pinto and Lambert, 2003). A sample ZCED is shown in Exhibit 3.

To establish the tolerance levels for voice-data switch users, define β_n as a real number between 0 and 1, inclusive such that

$$\beta_0 = 0 \text{ and } \beta_0 < \beta_1 < \beta_2 < \ldots < \beta_n. \quad (1)$$

If there are n numbers of different applications, each requiring a level of fraction of messages below a tolerance level β_n, then a set of configurations $f^*_n = \{f\}$ is said to be the appropriate set for application n if:

$$\beta_{n-1} \leq R^f_n(C_n) < \beta_n, \text{ for all f in the set } f^*_n. \quad (2)$$

Any departure from condition of Equation 2 is detrimental to the performance of a switch since. That is, $R^f_n(C_n) \geq \beta_n$ defines a switch is unreliable, and $R^f_n(C_n) < \beta_{n-1}$ defines a switch is underutilized. A switch user in the field of application n can choose among the configurations in the set f^*_n. As an example, consider Exhibit 3 showing three fields of applications: Safety-critical, business-critical, and non-critical identified as n=1, 2 and 3 respectively with tolerance levels $\beta_0 = 0$, $\beta_1 = 0.15$, $\beta_2 = 0.25$, and $\beta_3 = 1.0$. Then, for traffic rate T = 10000: $f^*_1 = \{5\}$, $f^*_2 = \{10,20\}$, and $f^*_3 = \{30,40,50\}$; for traffic rate T = 5000: $f^*_1 = \{5,10,20\}$, and $f^*_2 = \{30\}$, $f^*_3 = \{40,50\}$. The same process can be done for all traffic rates down to T = 100: for traffic rate T = 100: $f^*_1 = \{5,10,20,30\}$, $f^*_2 = \{40,50\}$, and $f^*_{n=3} = \{ \}$.

Exhibit 4. Sample ZCED.

Calls per hour						
10000	0.06	0.15	0.24	0.28	0.36	0.40
5000	0.06	0.13	0.14	0.23	0.26	0.28
1000	0.05	0.12	0.14	0.15	0.20	0.23
500	0.05	0.11	0.13	0.14	0.18	0.21
100	0.03	0.10	0.13	0.13	0.16	0.20
	5	10	20	30	40	50
	Number of large screen phones					

Exhibit 4 shows the resulting ZCED illustrating the different appropriate sets f^*_n using different shades of gray and labeled as zones 1, 2, and 3 corresponding to n = 1, 2, and 3. The chart also shows the different fractions of messages $R^f_n(C_n)$, where j=1,2..., n=1,2,3 and with a single threshold value 0.5 sec. for the LCD refresh function. The ZCED shows the configurations appropriate for a switch application with a certain traffic rate, and the tradeoff between switch configuration and performance in terms of latency times. The ZCED contains the important information for decision-making: criteria or performance measure relevant to the decision makers (Hillier and Lieberman 1986, and Pomerol and Barba-Romero 2000). In this case, these are the alternatives (f), and the payoff table or decision matrix ($Rf_n(C_n)$).

Super-Saturated PS

Special case is when the PS is super-saturated, i.e. when the rate of arrival of transactions is greater than the rate of departure – a realm of extreme value analysis. In application, this case will result to a build-up of transactions waiting for service and the corresponding increase in their waiting times. Thus, it is critical to know the necessary conditions when the waiting times of transactions in a saturated PS will increase with or without bounds (i.e. its distribution has bounded or unbounded tails).

In extreme value analysis, the interest is in the parametric model of the distribution of the maxima and the exceedances, which basically involves the tails of the distributions. The probability density functions and cumulative density functions of the maxima and minima of the latency times can be derived if the parent distribution and sample size of the latency times is known. Castillo (1987) and Castillo et al. (1989) has developed methods for determining the type of the extreme value distribution given a set of independent and identically distributed data. The different types of extreme value distribution are Gumbel Type I, which has an exponential tail, the Gumbel Type II, which has polynomial tail, and Gumbel Type III or bounded tail. The methods are based on the principle that distribution functions belonging to different Gumbel Types have different curvatures when drawn on Gumbel probability paper.

There can be difficulties in assessing the true characteristics of the tails of the distribution of the latency times for variations of PS. The difficulties arise from lack of knowledge on the cumulative probability function of the latency times due to incomplete information on the arrival and service of transactions, and on the possible variations in PS. This section presents a set of conditions that will assure that the tails of the distribution of latency times of classes in a PS are bounded or unbounded.

Proposition 1: Sufficiency conditions for unbounded latency times in PS. Sufficient conditions for unbounded transaction latency times in a saturated PS are:

i. the transaction service times are mutually independent,
ii. the transaction service times have common distribution.

Proof:

Consider the m*th* transaction to be served in the PS. The latency time of the transaction is composed of: remaining service time of the transaction being served when m arrived, assuming there is no preemption, service times of all transaction that will be served ahead of m, and service time of transaction m. Define the length of the remaining service time of the transactions being served when m arrived as x_0, and x_i, i=1,2,…,m-1, are the service times of transactions that are served ahead of m. The latency of transaction m is therefore the sum of these service times, that is

$$x_m = y_0 + y_1 + y_2 + \cdots + y_{m-1} + y_m$$
$$= \sum_{i=0}^{m} y_i$$

If the service times and the number of transactions (m-1) already in waiting for service are assumed to be random variables, then x_m , the sum of a series of random variables x_i is also a random variable. That is, if y_i is a random variable, i=0,1,2,…,m-1, and m is a random variate; then

$x_m = \sum_{i=0}^{m} y_i$ is also a random variable.

By the central limit theorem, x_m have a normal distribution function for large values of m. Since super-saturated PS is characterized by increasing number of waiting transactions, that is $m \to \infty$ as $t \to \infty$, the central limit theorem applies. The domain of attraction of a normal distribution is Gumbel type I (Castillo, 1988).

Note that Proposition 1 is applicable regardless of the priority class of the m^{th} transaction or the underlying distribution of the service times.

Proposition 2: Sufficiency conditions for bounded latency times in PS. For a transaction of a particular Class n in a PS, the sufficient conditions for bounded latency time are:

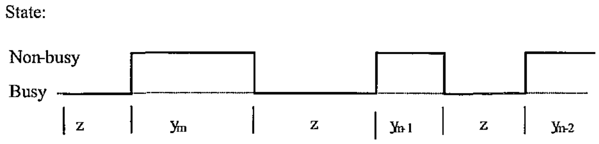
i. the service times of transactions for all classes are bounded,
ii. the waiting line for Class n is bounded, and
iii. the number of transactions with higher priority than Class n is finite.

Proof:

Consider an PS with N classes of transactions and with particular interest to Class *n*. The latency time of the m^{th} transaction of Class n is composed of: the remaining service time of the transaction being served when m arrived, assuming there is no preemption, service times of all transaction that will be served ahead of m, and service time of transaction m.

Consider two states of the server as busy and none-busy. A busy state is when the server is serving transactions belonging to class other than n, while a "non-busy" state is when the server is serving a transaction from Class n. Exhibit 5 illustrates the states where y_n represents the service time of transactions from Class n, and z represents the service times of transactions served in-between the service of transactions from the Class n.

Exhibit 5. Two-State Server in an PS defined by the Class of Transaction Being Served.



Therefore, the latency time of the m^{th} transaction of Class n is:

$$x_m = z_1 + y_1 + \cdots z_{m-1} + y_{m-1} + z_m + y_m$$
$$= \sum_{i=1}^{m} y_i + \sum_{j=1}^{m} z_j$$

Consider the possible properties of the service times y_i, z_i, and m (the number of terms in the sequence). The random variables can either be bounded or unbounded, and the number of terms m in the sequence can be also be bounded or unbounded. Exhibit 6 shows that only for the case of bounded service times y_i, z_i, and bounded m are the latency time x_m can be bounded.

Exhibit 6. Combinations of Bounded and Unbounded Service Times and Number of Transactions Ahead of transaction m.

y_i, z_i	m	x_m
Bounded	Bounded	Bounded
Unbounded	Bounded	Unbounded
Unbounded	Bounded	Unbounded
Unbounded	Unbounded	Unbounded
Bounded	Unbounded	Unbounded
Bounded	Unbounded	Unbounded
Unbounded	Unbounded	Unbounded
Unbounded	Bounded	Unbounded

Consider case when x_m is bounded. Such is the case if y_i, z_i, and m are bounded. A bounded m implies two conditions: that the waiting line of Class n is bounded; and that the number of transactions that can possibly be served ahead of the transaction of interest (i.e., has a higher priority than Class n) has an upper limit. A bounded y_i, and z_i implies that the service times for all classes are bounded.

The first condition of Proposition 2 is satisfied only if service times are known to be bounded or are artificially bounded by ejecting transactions that have been served for a long period of time. The second condition can be implemented by configuring the PS with a limit on the number of waiting elements for Class n such that all transactions arriving after the limit is reached are dropped or rerouted. The third condition is more difficult to satisfy since there is very little control over the arrival of transactions, and most PS applications have an infinitely many potential transactions. Systems that have a defined beginning and end of availability times like banks and offices can satisfy the third condition. However, for systems with an expected availability of almost 100% of the time like e-business servers, the number of potential transactions is unlimited.

Conclusion

Simulation modeling, coupled with an evaluator diagram can be effectively used in the configuration among users, features, and perceived reliability subject to extreme operation of switches. Tolerance level was critical in helping users identify the configuration that suits their requirements, especially during extremely high rate of call traffic. The necessary conditions for bounded and unbounded latency times of priority systems helped form strategies in designing future evolutions of priority systems. This approach can provide managers with insights to configuring and operating various types of systems during extreme loads.

Acknowledgements

This work was made possible through the support of Prof. James Lambert of the Center for Risk Management of Engineering Systems at the University of Virginia, and Ove Villadsen and Chuck Gross of Comdial Corporation.

References

Castillo, Enrique and Janos Galambos, and Jose Maria Sarabia. "The selection of the domain of attraction of an extreme value distribution from a set of data," *Extreme Value Theory*, J. Husler, and R.D. Reiss, eds. Springer-Verlag, (1989) pp. 181-190.

Castillo, Enrique. Extreme Value Theory in Engineering. San Diego, Academic Press, (1988).

Hillier, Frederick S. and Gerald J. Lieberman, Introduction to Operations Research, 4th ed. Holden-Day, Inc., (1986).

Johnson, Norman, Samuel Kotz, and N. Balakrishnan, Continuous Univariate Distributions, Vol. 2. New York: John Wiley & Sons, (1995).

Pinto, C. Ariel and James H. Lambert, "Configuration of Inter-Office Switch under Extreme Traffic with Zone Configuration Evaluator Diagram," *Reliability Engineering and System Safety*, Vol. 79, (2003) pp. 369-375.

Pinto, C. Ariel and James H. Lambert, "Extreme Events in the Configuration of Priority Systems," *Reliability Engineering and System Safety*, Vol. 76, (2002), pp. 265-271.

Pomerol, Jean-Charles and Sergio Barba-Romero. Multicriterion Decision in Management, Principles and Practice, Kluwer Academic Publishers, (2000).

About the Author(s)

C. Ariel Pinto is an Assistant Professor of Engineering Management and Systems Engineering at Old Dominion University. His works focus on risk management in engineered systems and systems engineering. He has worked at Carnegie Mellon University's Software Industry Center on software security and quality. He also worked at the Center for Risk Management of Engineering Systems at the University of Virginia on various projects with the US Army Corps of Engineers, Virginia Department of Transportation, and Comdial Corporation. He received his Ph.D. from the University of Virginia and his M.S. and B.S. from the University of the Philippines.