Availability and Cost Monitoring in Datacenters

Using Mean Cumulative Functions

David Trindade, Swami Nathan

Sun Microsystems Inc.

{david.trindade,swami.nathan} @sun.com

Keywords : Availability analysis, field data, mean cumulative function, repairable system, nonparametric

Abstract

Datacenters contain servers, network routers, switches and storage arrays to host the computing and networking facilities of organizations. The availability is routinely monitored on an ongoing basis by simply recording the percentage of uptime to total time. Uptime can be viewed as the ability of the application, machines, or the entire facility to provide some acceptable level of service. The percentage of uptime is typically defined over a time window such as a week or a month and plotted in calendar time. Sometimes the cumulative uptime percentage is used. However, a summary statistic like percent uptime can be a poor measure since it does not account for time dependence of failures and multicensoring inherent in the data. Furthermore, since availability is a function of both frequency of outages and outage duration, percent uptime does not distinguish between large numbers of small outages and a small number of large outages. This paper describes approaches based on extensions to Mean Cumulative Functions that have been used extensively at Sun Microsystems to analyze the reliability of repairable systems.

1. Introduction

Availability is of paramount importance to enterprises. The most common approach to monitoring availability is to count the percentage of uptime in a time window, e.g., week, month or quarter. This figure is reported on a periodic basis and is used to display the trend in availability for a group of machines. This single value is popular among management because it satisfies the *one number syndrome*. In addition to availability, some practitioners report a Mean Time Between Failure (MTBF) and a Mean Time To Repair/Recovery (MTTR). These are again summary statistics that hide information and are not optimal measures to monitor availability. The following examples illustrate how summary statistics can be inadequate and potentially misleading.

Number of outages	1	2	5	10	20
Average outage duration (minutes)	52.5	26.25	10.5	5.25	2.625
Total outage (minutes)	52.5	52.5	52.5	52.5	52.5
Availability (%)	99.99	99.99	99.99	99.99	99.99

Table 1: Different ways to achieve the same availability (1 year duration)

Availability is a function of the number of outages and the outage duration. Four 9's (that is, 99.99%) availability can be achieved by several combinations of outage frequency and duration. See Table 1. A single availability value does not reveal whether there was one large outage or several small outages. Two customers with the same availability may have completely different responses. Some customers do not mind small outages as long as large outages are avoided. For

other customers, each small outage can result in significant costs as personnel need to spend time reporting the outage, identifying the root cause, and scheduling planned downtime to remediate the problem. Hence, appropriate metrics for field availability monitoring should incorporate both outage frequency and duration visibility.

The common metric for representing outage frequency is an MTBF (where failure is an outage as defined by the customer). An MTBF is calculated by accumulating all the operating hours and dividing that figure by the number of outages. It is an extremely popular metric but is based on assumptions that are rarely checked in practice. A key assumption is that all failures come from a single population failure distribution and the system follows a renewal process [3]. Furthermore, the times between events are assumed independent and exponentially distributed with a constant rate of occurrence. These assumptions are often not satisfied in practice, and using MTBF without understanding the implications of the assumptions can result in drawing erroneous conclusions [1].
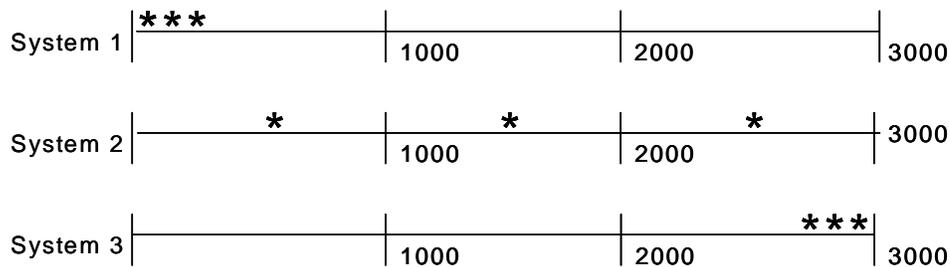
Figure 1: MTBF hides information

In Figure 1, we have 3 systems, each with 3000 hours of operation. All of them have three failures. The MTBF of all three systems is 1000 hours. System 1 has three failures early on and has experienced failure free operation since then. System 3 has three recent failures. System 2 has one failure in each 1000 hour time window. The system behaviors are quite different and yet

they have the same MTBF. One can similarly demonstrate the futility of using just an overall mean time to recovery (MTTR) to gain an understanding of outage durations.

In addition to these problems, data from populations of computer systems are subject to *multicensoring*. Since machines are installed throughout the year, at any point in time different machines have different ages, i.e., each machine is *right censored* at a different point. Similarly, historical failure data may not be available before a certain date, and hence there may be *left truncation* or *censoring* for various machines. The metrics discussed above merge multicensored data, e.g., it is impossible to ascertain from an MTBF if failures are occurring because old machines are getting worse or new machines are having early failures since these effects tend to be hidden by combining all failures into a single count.

We discourage the indiscriminate use of summary statistics like percent uptime, MTBF, and MTTR in monitoring availability in the field and instead present alternate techniques. These methods are based on extensions to Mean Cumulative Functions (MCF), which is a non-parametric approach that is increasingly being used to monitor the reliability of repairable systems in the field [2,3,4].

2. Mean Cumulative Functions

When analyzing repairable systems the simplest plot that can be constructed is a cumulative plot, which graphs the number of failures (outages) versus time, where time can be age from installation or calendar time. When we have a group of systems, each machine can be represented by a cumulative plot. It is also possible to represent the behavior of the group of machines by an average number of failures versus time. This average is known as the Mean Cumulative Function (MCF).
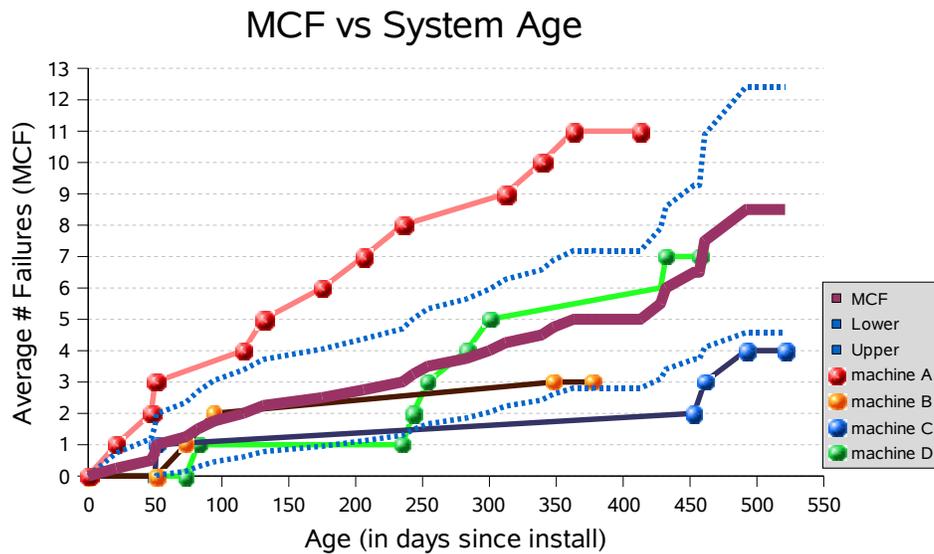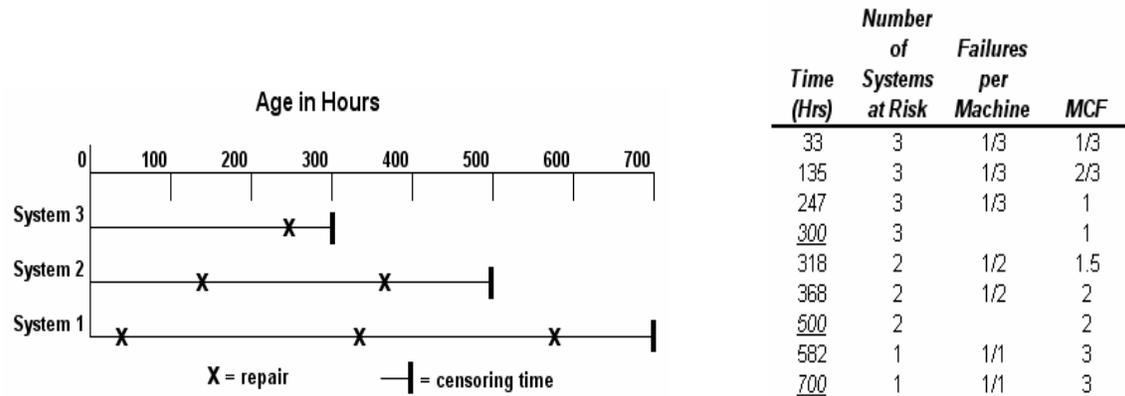
Figure 2: MCF and confidence bounds for a population of four machines

In Figure 2 we see the cumulative plots for four machines. The MCF is the average number of failures for this population as a function of the age of the system. Since different machines have different ages the MCF is calculated by normalizing the number of failures in a time window by the number of machines at risk. This method accounts for multicensoring. For example, a machine that is only 200 days old cannot contribute reliability information at 400 days (right censored). Similarly if data collection began on a particular machine only after 150 days, that machine cannot contribute information to the reliability at 75 days (left truncated or censored).

An example illustrating a step by step calculation of the MCF is shown below. The figure on the left shows the history on three systems. At the time of the analysis, System 1 has operated for 300 hours, System 2 for 500 hours, and System 3 for 700 hours. These represent censoring times, i.e., the systems cannot contribute information beyond their right censoring times. At 33 hours System 1 has a failure, and since there are three machines in the population at this age we get 1/3

fails per machine. The MCF aggregates the fails per machine at all points in time. At 318 hours, there is another failure but since there are only two machines in the population, the fails per machine is 1/2. The calculation of the MCF can be easily performed in a spreadsheet environment by following the example table.

Time (Hrs)	Number of Systems at Risk	Failures per Machine	MCF
33	3	1/3	1/3
135	3	1/3	2/3
247	3	1/3	1
300	3		1
318	2	1/2	1.5
368	2	1/2	2
500	2		2
582	1	1/1	3
700	1	1/1	3

Figure 3 : Step by Step MCF calculation.

Since the MCF is an estimate, confidence bounds can be calculated [2]. In Figure 2 one can see that the confidence bounds are narrower in the earlier ages where there is no censoring, while the bounds become wider at later ages where the population at risk diminishes due to right censoring.

The MCF versus age curve can be numerically differentiated to obtain the slope, called the recurrence rate (RR). This rate is identical to the *rate of occurrence of failures or ROCOF* found in repairable systems literature [5]. The rate tends to reveal trends in the MCF. If the MCF rises quickly, a sharp spike in the recurrence rate results, while a constant rate implies a homogenous Poisson process [3], where an MTBF has meaning.
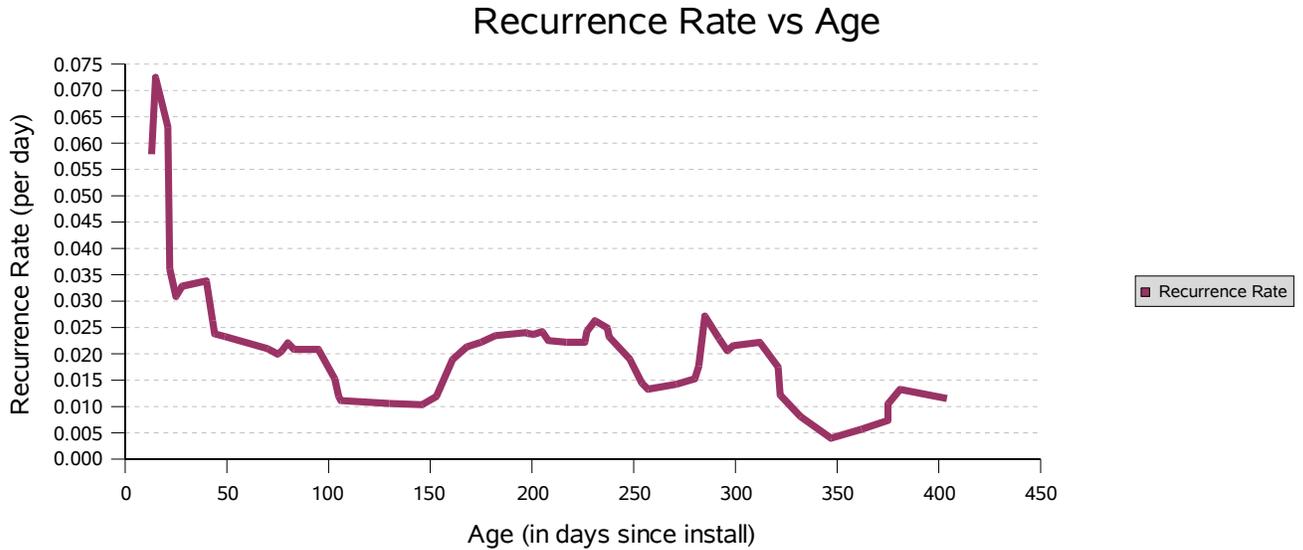
Figure 4: Recurrence rate versus age (slope of the MCF curve)

Figure 4 shows an example recurrence rate versus age for a population of machines. We can see that there were significant early life problems, and then the rate of failures steadily decreased, i.e., the reliability of this population of machines improved with age.

MCFs are a superior way for representing outage frequency compared to MTBFs for the following reasons

1. No distributional assumptions (nonparametric)
2. Explicitly accounts for right and left censoring
3. Provides trends in the rate of failures (outages) as a function of time.

In the next section we extend the notion of MCFs from just counts type data to continuous data to handle outage duration, the other component of availability.

3. Mean Cumulative Downtime Functions

In addition to describing failure count data, MCFs can be easily extended to handle continuous

data such as downtime, cost, power utilized etc.

Machine	Age	Event	Duration (mins)
1	23	outage	60
1	104	outage	25
1	300	outage	5
1	400	censor	-
2	120	outage	30
2	460	outage	10
2	500	censor	-
3	75	outage	2
3	80	outage	2
3	85	outage	2
3	90	outage	2
3	95	outage	20
3	700	outage	5
3	800	censor	-

Table 2 : Example outage data on 3 machines

In Table 2 we have the outage history on three machines. Machine 1 is 400 days old (censoring time) and has three outages of 60, 25 and 5 minutes. Machine 2 is 500 days old, and Machine 3 has been operational for 800 days.
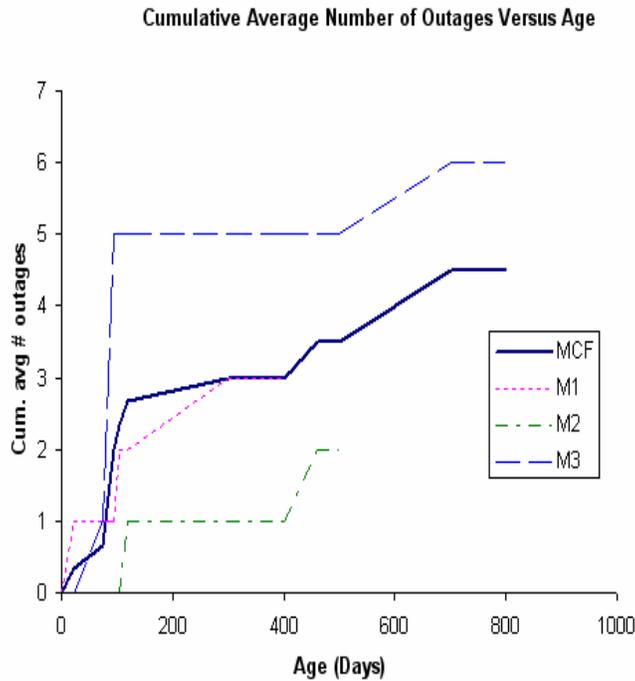

Figure 5: Cumulative Plots and MCF for the 3 machines.

Following the procedure outlined in Figure 3, the MCF for these three machines based on outages is calculated. One sees in Figure 5 that M3 has experienced many outages in a very short time period. After the problem was remediated, there were long periods of failure free operation. Such clustering of events, typically indicate incorrect diagnosis or inability to fix the problem on the first try. In this example the sample size is small but when we have a sizeable number of machines under observation, the MCF gives one an excellent idea of how many outages to expect in every age window. The shape of the curve provides clues to the existence of early life issues

or wearout mechanisms. However, the MCF doesn't indicate how many minutes of downtime have occurred. In fact, we reasonably assume in creating such plots that downtime can be neglected compared to the lifetimes of systems. The MCF reveals one facet of availability, i.e., outage frequency.

The same procedure in Table 3 can be used to calculate a mean cumulative downtime function by averaging the individual cumulative downtime plots which show how much downtime has been accumulated by each machine. In Figure 6, we see that M1 has had 90 minutes of downtime in 400 days. Although M3 had 5 outages within 100 days of operation (figure 5), we see in Figure 6 that the amount of downtime is much less than M1's downtime. This observation follows because M1 had a single event that resulted in 60 minutes of downtime. M3's events consist of four outages of two minutes each followed by a 20 minute outage. This situation often arises in the real world where we have a restart after a transient error, and after a few such iterations, we schedule a longer outage to fix the problem to avoid a non-transient failure.
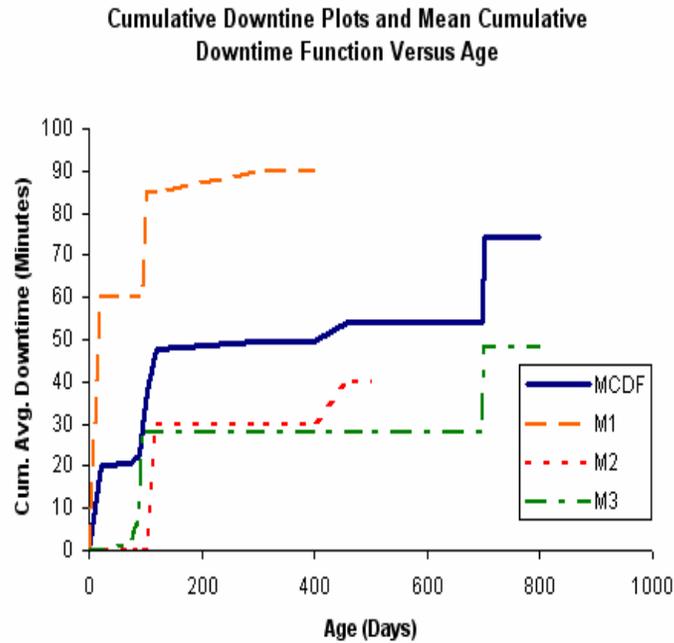
Figure 6: Cumulative downtime plots and mean cumulative downtime function

Similar to MCFs, one can estimate recurrence rates via the slope to determine if the rate of accumulation of downtime per machine is increasing or decreasing for this population.

4. Calendar Time Analysis

MCFs and mean cumulative downtime functions (MCDFs) as a function of age reveal age related effects such as early life failures or wear out mechanisms. Often effects related to calendar time play an important role in the availability of datacenter systems, e.g., patches, upgrades to new versions of software or faster cpus, relocation of machines to new datacenters, etc. The same analysis can be performed as a function of calendar time and it can be quite revealing. Details can be found in [1]. Customers and service personnel relate easily to calendar time MCFs which can point to specific actions that occurred in the datacenter during that week or month.

5. Time Dependent Failure/outage Cause Plots

The common approach to viewing the failure causes is by means of a Pareto chart or bar chart. These charts are static and do not reveal which failure modes are a current threat and which ones have been remediated. One can construct plots based on counts of failure causes or mean cumulative downtime failure cause functions by segregating the data into outages related to a specific cause and calculating the MCF. These can again be plotted as a function of age and calendar time.

6. Comparisons of MCFs

One of the most useful features of MCFs and extensions to continuous variables is that MCFs provide the ability to make statistically meaningful comparisons between multiple populations. Comparing the availability or MTBF or MTTR of two populations does not provide an adequate comparison because these measures are not compared at all points in time, only at a gross average level. With MCFs and MCDFs one can compare a population of machines with another at all points in time making for a more valid and informative comparison. One can compare similar application solutions across different customers, compare one datacenter with another for the same customer, compare machines from one vintage with machines from another vintage, machines in production with machines in development and test and so on.
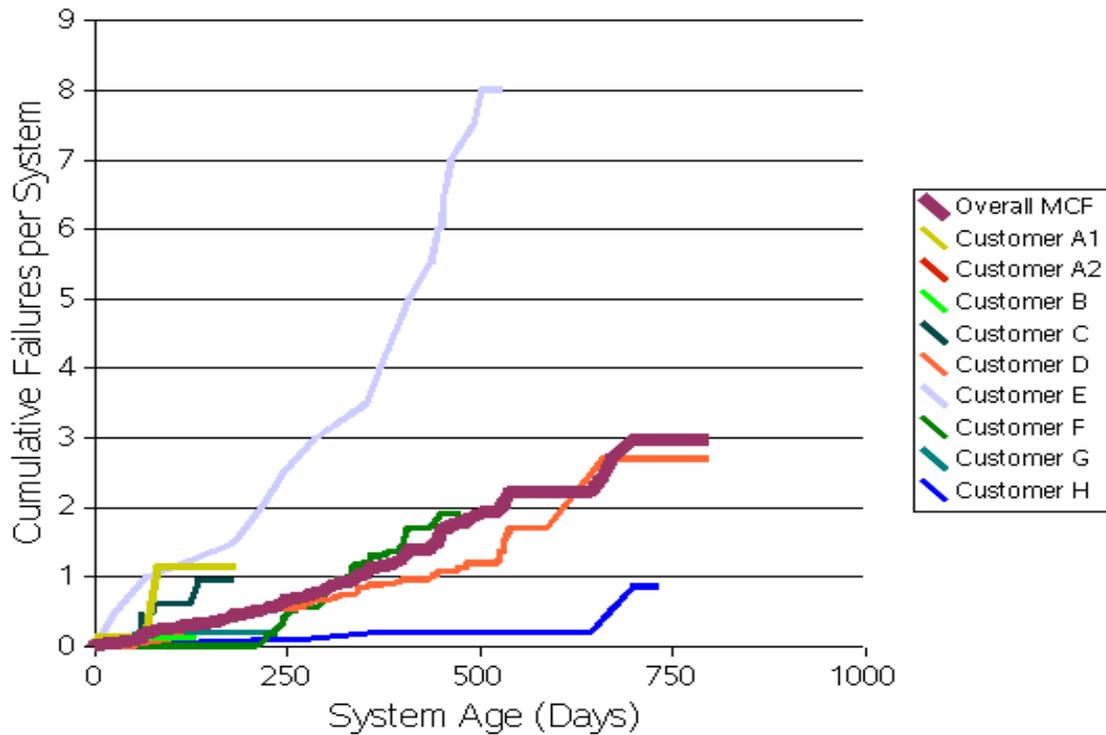
Figure 7 : Comparing MCFs across customers

Figure 7, shows a comparison of MCFs across several customers. One sees that most are performing similarly, but Customer E has been experiencing a significantly higher number of failures than all the other customers at all ages. Further investigations at this customer revealed process issues. One can similarly compare cumulative downtime curves across customers running similar applications and determine if any population is performing beyond the norms of statistical variation. Such anomalous populations can be targeted for investigations to understand the nature of such variation, which can be a source of tremendous learning.

7. Mean Cumulative Cost Function

Another important extension to MCF is to calculate the mean cumulative cost. The process is identical to calculating the mean cumulative downtime function except downtime is replaced by

cost. We can incur a cost per minute of downtime due to lack of services or loss of revenue. We can also realize a cost per outage event because of contacting support services and associated failure analysis reviews, etc. Often designers create high availability solutions by clustering and redundancy and can guarantee low downtime, but total cost of operation could be high because of outage frequency. While architecting solutions, care should be taken to optimize based on total cost and not strictly on amount of downtime. In the example data in Table 2, assume that each outage event incurs 200 units of cost and each minute of downtime incurs 10 units of cost. We can now plot mean cumulative cost functions for event, downtime, and total costs, for individual machines as well for groups of systems.
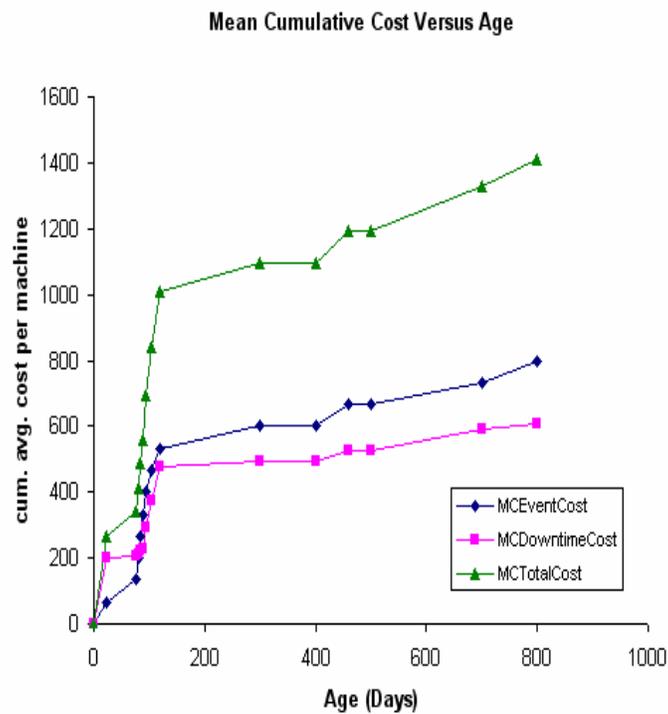
Figure 8: Mean cumulative cost function versus age

In Figure 8, we plot the mean cumulative cost functions. We can plot total cost, downtime cost or event cost/repair cost. When we have such plots based on a sizeable population of machines, we

can obtain an excellent idea of the structure of the total cost of operation for a platform or a solution. Given this information, one can design service level agreements (SLAs) in terms of outage frequency or downtime and design innovative award and penalty schemes to assure compliance with the guarantees. One can be quite confident in pricing these SLAs because we would have a good handle on the cost structure as a function of time.

8. Conclusions

Analyzing and monitoring availability based on field data often has been limited to calculating summary statistics. Modeling outage frequency and cumulative outage duration as a function of time is far more revealing than the "*number of nines*". By using techniques based on mean cumulative functions, a much richer analysis can be created from the same data. These functions are distribution free, can be easily implemented in a spreadsheet environment, and are understandable by managers and laymen. The mean cumulative cost and downtime functions can be used effectively to create service level guarantees, set optimal prices for various guarantees, and establish award and penalty schemes for compliance with guarantees.

References

1. D.C. Trindade, Swami Nathan, "Simple Plots for Monitoring the Field Reliability of Repairable Systems", *Proceedings of the Annual Reliability and Maintainability Symposium (RAMS)*, Alexandria, Jan 2005.

2. W. Nelson, *Recurrence Events Data Analysis for Product Repairs, Disease Recurrence and Other Applications*, ASA-SIAM Series in Statistics and Applied Probability, 2003.

3. P.A. Tobias, D.C. Trindade, *Applied Reliability*, 2nd ed., Chapman and Hall/CRC, 1995.

4. W.Q. Meeker, L.A. Escobar, *Statistical Methods for Reliability Data,* Wiley Interscience, 1998.

5. H. Ascher, H. Feingold, *Repairable Systems Reliability: Modeling, Inference, Misconceptons and their Causes*, Marcel Dekker, 1984.