Optimal IT Service Availability: Shorter Outages, or Fewer?

Ulrik Franke

***Abstract*—High enterprise IT service availability is a key success factor throughout many industries. While understanding of the economic importance of availability management is becoming more widespread, the implications for management of Service Level Agreements (SLAs) and thinking about availability risk management are just beginning to unfold.**

This paper offers a framework within which to think about availability management, highlighting the importance of variance of outage costs. The importance of variance is demonstrated using simulations on existing data sets of revenue data. An important implication is that when outage costs are proportional to outage duration, more but shorter outages should be preferred to fewer but longer, in order to minimize variance.

Furthermore, two archetypal cases where the cost of an outage depends non-linearly on its duration are considered. An optimal outage length is derived, and some guidance is also given for its application when the variance of hourly downtime costs is considered.

The paper is concluded with a discussion about the feasibility of the method, its practitioner relevance and its implications for SLA management.

***Index Terms*—Service level agreements, optimization techniques, availability, SLA management, fault management, policy-based management.**

I. Introduction

MANAGING enterprise IT services in order to maintain high availability is a difficult yet very important task. In a classic 1998 survey, IBM Global Services reported that the cost of unplanned downtime to American businesses was $4.54 billion in 1996 [1]. The situation has not improved since. A Gartner report, based on surveys conducted in 2007 and 2008, concludes that the ”overall proportion of mission-critical IT services continues to increase, along with the cost of business downtime” [2]. In a recent survey, 178 enterprise IT system executives and practitioners from Sweden and the German-speaking countries were asked to assess future IT system quality priorities in their companies. On a five point Likert scale, 48.9% of respondents gave availability the highest mark, making it the most highly prioritized system quality in the survey [3]. Conversely, reliability and its costs were identified as the second highest frustration with IT in a 2010 survey among CEOs and senior business executives [4]. Furthermore, the market value of publicly traded companies is affected by IT incidents which disrupt business operations [5], [6], even though this effect might be small [7].

Manuscript received February 22, 2011; revised July 20, 2011 and October 13, 2011. The associate editor coordinating the review of this paper and approving it for publication was M. Brunner.

U. Franke is with the Department of Industrial Information and Control Systems, Royal Institute of Technology (KTH), 100 44 Stockholm, Sweden (e-mail: ulrikf@ics.kth.se).

Digital Object Identifier 10.1109/TNSM.2011.110811.110122

Design of high availability systems is an old field, but over the decades focus has changed. Fifty years ago, hardware components were the major source of faults and outages, whereas operations, environment, and software faults have since caught up [8]. This underscores the importance of *business* continuity [9], [10], where the focus is on the services delivered, not on the systems providing them. Naturally, executive visibility of the costs associated with IT service downtime is important [11], as is management of service level agreements (SLAs) with availability specifications. One source, among many, giving practical recommendations on SLAs is [12].

This paper offers a framework for thinking about IT service availability that is more mature than the simple percentage concept (e.g. 99.9%). Doing so, we will address at length the issue of variance and the role it plays in investments and SLA management. In a nutshell, the paper seeks to establish a view of IT service availability similar to the mainstream view on capital investment: any strategy for maximizing expected returns must be considered also in the light of its risk, i.e. the variance of the expected returns.

A. Outline

The remainder of the paper is structured as follows: Section II introduces some related work. Section III introduces a basic model of availability investment, which is extended in Section IV. Some empirical examples follow in Section V. Different decision-making strategies with regard to availability are discussed in Section VI. Section VII discusses the feasibility of applying the analytical model in practice and its relevance for practitioners. Finally, Section VIII summarizes the contribution and offers some concluding remarks.

II. Related Work

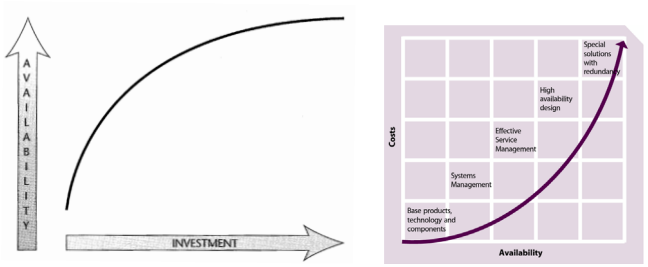
Some good textbooks on the general area of IT systems availability are [13], [14]. In [15] the authors present an approach for analytical service availability assessment, mapping dependencies between low-level ICT infrastructure and high-level services. In [16] a similar mapping is presented, but here the focus is on the impact of ICT infrastructure availability upon business processes. An effort to identify factors impacting software reliability is presented in [17], where 32 factors impacting software reliability are identified. In [18], 16 factors affecting enterprise IT systems availability are presented, subsequently evaluated in 9 case studies [19]. However, most work on availability is delimited to very particular technical areas. End-to-end discussions on how IT systems availability impact business revenue are quite rare. While in [18], [19]

the whole enterprise architecture is considered, the percentage concept of availability is retained.

However, an early consideration of the market for "faultless" telecommunications introduces an outage-cost concept similar to ours [20]. The argument about telecommunications can be readily extended also to more modern ICT systems. Another work closely related to the present article is [21], where the impact of computer breakdowns on pricing and capacity decisions in companies is examined. One result is that backup capacity should be acquired when computing is critical to the firm. However, no analysis is made of the optimal use of these backups. Another article by the same author investigates pricing and capacity decisions of clustered twin-computer systems subject to breakdowns [22]. In this context, the conclusion is that it pays to reduce the mean time to repair rather than to increase the mean time between failures. This is an issue at the heart of the present article, though we consider it in a more general setting.

One well-established and mature area of economic analysis of the worth of reliability is power systems economics. Even though reliability or availability of power delivery is not identical to the issue of IT systems, parallels can be drawn. This is all the more true in light of recent arguments about the "commoditization of IT", making it more like electricity [23]. For example, in [24] the reliability of power distribution systems is investigated from a cost perspective: redundancy schemes improve system reliability, but also entail double investments. In [25], a cost-benefit method for determining optimal power system capacity and reliability of electricity supply using customer shortage costs is described. There is also work, with numerical results, on finding reliability levels that match customers' willingness to pay [26]. An interesting study on the impact of electricity outages on industrial customers is reported in [27], indicating a large variation in customer preferences for reliability. In [28], the willingness to pay for reductions in power outages is investigated. The results indicate that the marginal willingness to pay increases with the duration of the previous outages and is higher if previous outages occur during weekends and during winter months. This highlights the importance of considering availability not only as a mere percentage of uptime, but also its distribution over time. Another supply security study is [29], where a trade-off perspective for finding the optimal level is adopted: reductions in the number of power interruptions are often possible, but also very costly.

The research area of optimal SLA management is still growing, but has produced some interesting work. In [30] decision models are derived that aim for optimally scheduled outages (service-windows) in terms of lowest expected business impact. This is similar to the present paper, except that our interest is *un*-scheduled outages. Other research such as [31] is related insofar as it attempts to find optimal managing and pricing strategies for SLAs, but different in that availability is not in focus. Still other papers such as [32] consider first and foremost the specification of SLAs, rather than quantitative risk analysis of them.



(a) Marcus & Stern version (b) ITIL version

Fig. 1. The availability continuum graphed against investment, reprinted from [13] and the same with the axes flipped, reprinted from [34]. Note that no scales are specified in either graph.

III. The Decision-maker's Problem

Availability refers to the ability of an item to perform its required function at a stated instant of time or over a stated period of time [33]. The *average availability* can be computed as the Mean Time To Failure (MTTF) divided with the total time of operation, which is the sum of MTTF and the Mean Time To Repair or Mean Time To Restore (MTTR) as seen in Eq. (1) [33]:

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}} \tag{1}$$

A more cautious availability estimate is found by instead using the "Maximum Time To Repair", corresponding to a worst-case scenario.

In a modern service-oriented environment of enterprise information systems, availability is most often addressed at the management level by contractual instruments such as Service Level Agreements (SLA), Operational Level Agreements (OLA), and Underpinning Contracts (UC), to use the terminology of ITIL [34]. This raises the question of how these documents ought best to be written. This question is both qualitative – what KPIs should be used? – and quantitative – what are their appropriate levels? The aim of the following discussion is to offer a framework within which to think about these issues.

A conceptual view of the relation between the investment in availability of IT systems and the resulting actual availability, reprinted in Fig. 1(a), is offered in [13]. A similar graph is given in the popular ITIL practitioners' framework [34], seen in Fig. 1(b). What is displayed is a classic example of diminishing returns. As described in [13], "costs increase until the cost curve nearly flattens out along the top" and "the highest levels of availability approach but never quite reach 100 percent". The ITIL text similarly speaks of "the exponential cost of delivering higher levels of availability". Formalizing these descriptions, a function $A = f(A_0, c)$ describing the relationship between investment $c \geq 0$ (c for cost) and resulting availability $A \in [0, 1]$ should meet the following requirements:

- R1. $f(A_0, 0) = A_0$. This means that we start out at a reference level A_0, and if no investment is made, we remain there.
- R2. $f_c(A_0, c) > 0$. The function is monotonously increasing, so any investment results in higher availability.

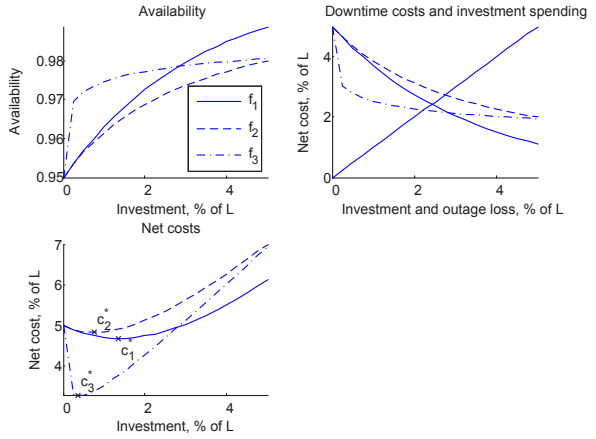


Fig. 2. Availability (upper left), separate downtime costs and investment costs (upper rights), and net costs according to Eq. 3 (lower left). $\alpha = 0.3$ for all functions.

R3. $f_{cc}(A_0, c) < 0$. The rate of increase, however, is diminishing, so the availability improvement achieved for each additional dollar invested shrinks.

R4. $\lim_{c\to\infty} f(A_0, c) = 1$. As the amount invested in availability approaches infinity, availability does approach 100%.

Cases where investment leads to increased technical complexity – and thus to more human errors – challenge monotonicity. However, investments in availability should be broadly construed to include both technical and organizational investments, e.g. staff training.

What is gained by increasing availability? A simple formula for the estimated average cost of 1 hour of downtime is the following [35]:

$$\begin{array}{ll} & \text{Empl. costs/hour} \cdot \,\%\text{ Empl's affected by outage} \\ + & \text{Avg. Rev./hour} \cdot \,\%\text{ Rev. affected by outage} \\ \hline = & \text{Estimated average cost of 1 hour of downtime} \end{array} \quad (2)$$

Multiplying this cost estimate with the number of hours per operating year (e.g. 365 days · 24 hours for 24-7 systems), gives a maximum potential loss L, a percentage of which is actually lost due to unavailability. With availability A, the losses are $(1 - A)L$, e.g. $A = 95\%$ entails a loss of $0.05L$. Observe that this hourly cost is constant, regardless of outage duration, an assumption to be further analyzed in Section VI.

Adding the downtime costs to the investment costs, we obtain a net cost function:

$$\text{Net cost} = (1 - f(A_0, c))L + c \quad (3)$$

The rational decision maker ought to find the level of investment c^* that minimizes this function. Note that c^* will always be smaller than $(1 - A_0)L$, i.e. smaller than the expected costs from outages ex ante. To see this, observe that $0 < A_0L - f(A_0, c^*)L - c^* < A_0L - c^*$, so $0 < A_0L - c^*$. Thus, it is *not* the case that the budget for availability improvements should equal the expected costs of outages (this observation echoes one made in [36] on investments in IT security). Figure 2 gives some graphical examples of different functions that meet the requirements R1-R4:

$$\begin{array}{rcl} f_1(A_0, c) & = & 1 - (1 - A_0)e^{-\alpha c} \\ f_2(A_0, c) & = & 1 - \frac{1-A_0}{1+\alpha c} \\ f_3(A_0, c) & = & A_0^{\frac{1}{1+c^\alpha}} \end{array} \quad (4)$$

All the functions are parametrized with a parameter $\alpha \in (0, 1)$, affecting their shape. It is, of course, difficult to see whether function 1, 2, or 3 is the most likely candidate for the function in Fig. 1(a), as depicted by [13] – or if still another function is more suitable. Finding a suitable function is not our purpose here – rather we discuss the issue conceptually.

Functions $f_1 - f_3$ are depicted in Fig. 1(a), with $A_0 = .95\%$ and $\alpha = 0.3$, to give an impression of their characteristics. In the upper right sub-figure, the resulting availabilities have been converted into downtime costs (i.e. $1 - f_n(A_0, c)L$), and supplemented with the investment cost c (i.e. the $x = y$ line). This amounts to graphing the two components of Eq. 3 separately, before adding them together as in the lower left sub-figure.

To minimize the net cost, we need to find the zero of the derivative of Eq. 3 with respect to c (assuming that the optimal investment level c^* is an interior point). This first order condition on c^* is:

$$f_c(A_0, c^*)L = 1 \quad (5)$$

For functions f_1 and f_2 this readily translates into nice closed-form optima, whereas the f_3 condition is somewhat more cumbersome:

$$\begin{array}{rcl} c_1^* & = & \frac{\ln(\alpha \cdot L \cdot (1-A_0))}{\alpha} \\ c_2^* & = & \frac{\sqrt{(1-A_0)\cdot\alpha\cdot L} - 1}{\alpha} \\ & & \frac{\alpha c_3^{*\alpha-1}}{(1+c_3^{*\alpha})^2} = -A_0^{1+c_3^{*\alpha}} L \ln A_0 \text{ must hold for } c = c_3^* \end{array} \quad (6)$$

These optima are marked to the lower left in Fig. 2.

The model outlined so far is clearly an idealized one. Two important complications can be readily discerned:

1) The use of *average* hourly revenues. The variance of hourly revenues might make such an average misleading.
2) The use of average *hourly* downtime costs. Downtime costs can vary with outage duration.

The focus of this article is the first of these complications, which is addressed in the subsequent sections. Looking at Eq. 2, the hourly revenues seem most prone to exhibit a large variance. The second complication is then reviewed in the latter part of Section VI.

IV. The Variance of Average Revenues

What happens to this optimization approach if the hourly costs of downtime can vary? The simplest way to reflect the variability of downtime costs is to make L a stochastic variable. The stochastic version of Eq. 3 uses the expectation of L:

$$\text{Net cost} = (1 - f(A_0, c))E[L] + c \quad (7)$$

A confidence interval is found by adding or subtracting n (not necessarily an integer) standard deviations from the result:

$$\text{Net cost} = (1 - f(A_0, c))(E[L] \pm n\sqrt{V[L]}) + c \quad (8)$$

For an unknown distribution of L, the Chebyshev inequality can be used:

$$P(|L-\mu| \geq n\sigma) \leq \frac{1}{n^2} \tag{9}$$

μ is the expectation and σ is the standard deviation of L. Thus, at worst, 50% of observations will fall outside of $E[L] \pm \sqrt{2}\sqrt{V[L]}$, decreasing to 25% for $n = 2$, 11% for $n = 3$ etc. For some known distributions we are much better off.

However, a better model accounts for the fact that the time of occurrence greatly influences the outage cost. The total cost of downtime is a sum over the set Out of hours when outages occur, where each hour has a separate random cost variable L_i. The expected total cost is:

$$\text{Net cost} = (1 - f(A_0, c)) \sum_{i \in Out} E[L_i] + c \tag{10}$$

The variances are more complicated. In general, $V[X + Y] = V[X]+V[Y]+2Cov[X, Y]$. There are two extreme cases to this equation: For *independent* random variables X and Y, $V[X+Y] = V[X]+V[Y]$, since the covariance is zero. For a sum of the *same* random variable $V[X + X] = 4V[X]$, since $Cov[X, X] = V[X]$, and in general $V[aX] = a^2V[X]$. This means that the variance exhibited by downtime costs summed over outage hours can take several (approximate) forms:

$$V\left[\sum_{i \in Out} L_i\right] \begin{cases} \approx & |Out|^2 \cdot V[L_i] & \text{all } L_i \text{ approximately the same} \\ = & \sum_{i \in Out} V[L_i] + 2\sum\sum_{i<j} Cov[L_i, L_j] & \text{in general} \\ \approx & \sum_{i \in Out} V[L_i] & \text{all } L_i \text{ approximately independent} \end{cases} \tag{11}$$

Are any of the approximate cases plausible? One way to look at the problem is as follows: assume that the random variables in the series of hours for the whole year (365 days times 24 hours for continuously running systems) $L_1, \ldots, L_{365 \cdot 24}$ all belong to some distribution D: $L_i \in D$. If outage hours are picked uniformly over the year, the $L_i \in Out$ are more or less independent, and the variance of their sum should look more or less like the third alternative above. However, if outage hours are not uniformly picked, but appear together in longer outages of consecutive hours, they become more dependent. For example, two consecutive outage hours in a retail business before Christmas probably have a greater covariance than one hour from before Christmas and one hour from a February Monday morning. So, as outages become fewer, we move towards the first case – all L_i approximately the same – which might be a useful approximation for a single long outage of consecutive (and similar) hours.

Thus, the number of outages plays an important role for the variance of downtime costs, since it determines whether outage hours are mostly consecutive or non-consecutive. The most basic model used in reliability theory to find this number is to model the occurrence of failures as a homogeneous Poisson process (HPP) [37], which will be used in this paper. The probability that a failure occurs exactly n times in the time interval $[0, t]$ is

$$P(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \text{ for } n \in \mathbb{N} \tag{12}$$

$N(t)$ belongs to the Poisson distribution: $N(t) \in Po(\lambda t)$. λt is the expected number of outages in $[0, t]$: the product of λ, the intensity of the HPP [occurrences/time] and t, the length of the time interval.

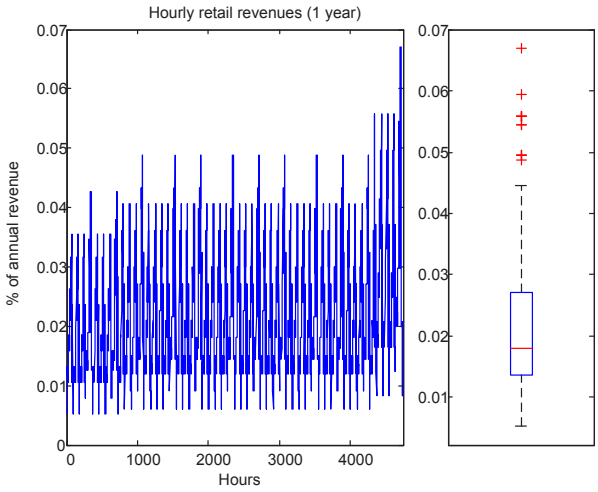


Fig. 3. Overview of Dataset 1.

V. Empirical Revenue Data

To shed some empirical light on the theoretical discussion of the previous section, this section applies the theory to two datasets of approximate downtime costs.

A. Dataset 1

The first dataset reflects hourly revenues for the Swedish retail business during a year. It is based on [38], a report from the Swedish Retail Institute (HUI) with revenue statistics from the Swedish retail sector. The report contains annual revenue distribution over months, weekly revenue distributions over hours from 9 a.m. to 10 p.m. each day (for normal weeks and payweeks) and the impact of holidays (e.g. Christmas and Easter). Based on these statistics, a dataset of 13 hours times 365 days was generated and normalized, reflecting relative hourly revenues over the year. The dataset most closely reflects revenues in the consumer goods market, since the hourly statistics is based on supermarkets. An overview of Dataset 1 is given in Fig. 3, which also presents a boxplot with the lower quartile, median, and upper quartile values marked. As seen in Fig. 3, there is lot of variance. The hour with the least revenue contributes 0.0053% of annual revenue, whereas the hour with the most revenue contributes 0.0670%, i.e. more than 12 times as much. A single high revenue hour outage might cost as much as a dozen low revenue hour outages. Furthermore, since the dataset has been computed backwards from statistical averages, this variance is actually a lower bound.

Eq. 11 requires a suitable model for L_i, the stochastic variable representing hourly downtime costs. Using the standard method of quantile-quantile plots, we find that Dataset 1 can be approximated reasonably well by a Gamma distribution, a Rayleigh distribution or a Weibull distribution. For its mathematical simplicity, we use the Rayleigh distribution, $L_i \in R(\sigma)$:

$$f_{L_i}(x) = \tfrac{x}{\sigma^2} e^{-x^2/2\sigma^2} \quad E[L_i] = \sigma\sqrt{\pi/2} \quad V[L_i] = 2\sigma^2\left(1 - \tfrac{\pi}{4}\right)$$

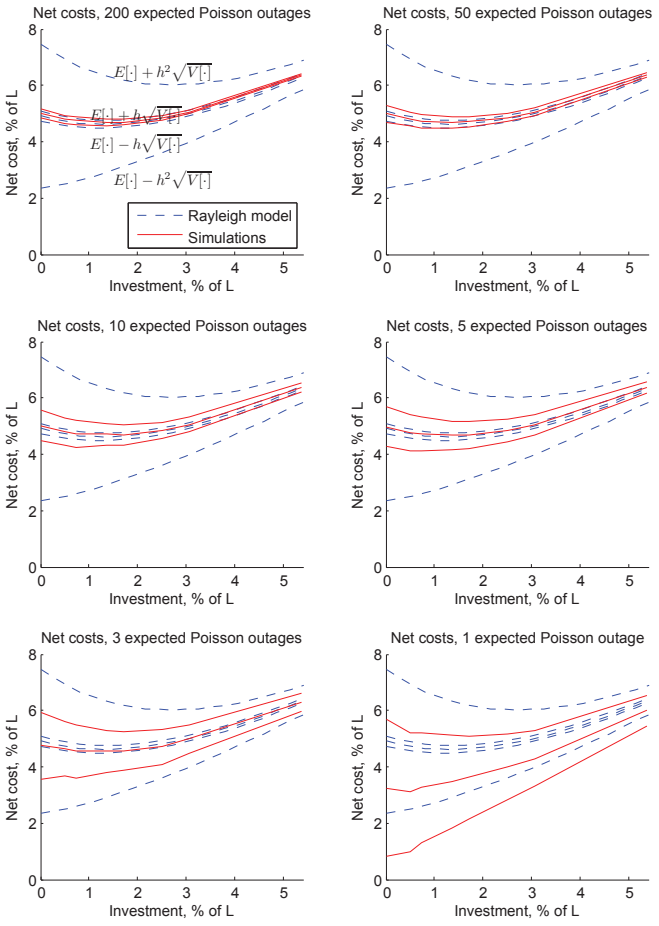


Fig. 4. Means and standard deviations as predicted by the Rayleigh model and as simulated from Dataset 1.

For Dataset 1, the maximum likelihood estimate gives $\sigma \approx 0.0165$.

To investigate the implications for optimal decision-making in terms of Eqs. 10 and 11, Monte Carlo simulations were made where outage hours were randomly selected from Dataset 1 corresponding (i) to different availability investments, giving different availabilities according to f_1 of Fig. 2, (ii) to different expected numbers of outages according to Eq. 12. For example, the investment case resulting in 97% availability corresponds to simulating $3\% \cdot 4745 = 142.35$ hours of downtime. These hours were drawn from Dataset 1 with an expected number n of 200, 50, 10, 5, 3, and 1 outages by Eq. 12, i.e. with expected average outage length ranging from less than one hour to more than 142 hours. Each of these cases was simulated 1 000 times. The results are plotted in Fig. 4. It should be stressed that the six scenarios all show the same levels of availability – the difference is that the downtime is distributed over more or less outages. In each of the sub-figures in Fig. 4, as we move to the right the investment in availability is increased, as is availability as such. In this respect, Fig. 4 is just like the lower left sub-figure of Fig. 2. In each of the six sub-figures, the blue dashed lines are depictions of the Rayleigh model. Five such lines are plotted: the one in the middle is the mean (i.e. $E[L_i] = \sigma\sqrt{\pi/2}$), moving out from the middle it is followed by two lines denoting a single standard deviation ($E[L_i] \pm h\sqrt{V[L_i]}$) under the assumption

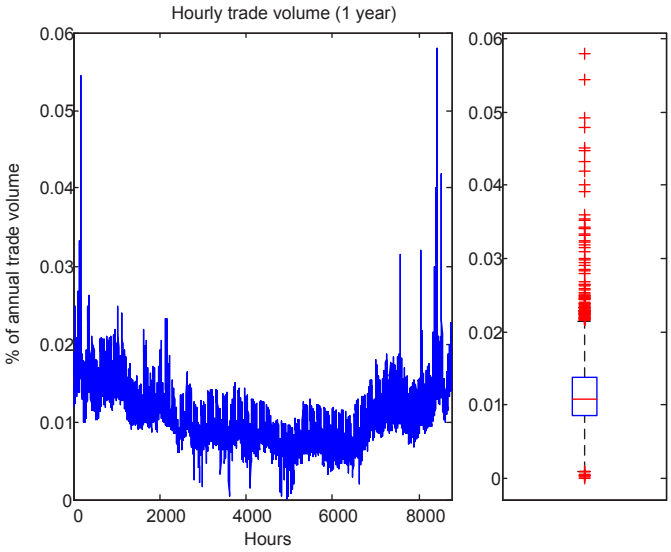


Fig. 5. Overview of Dataset 2.

of independent L_i (cf. Eq. 11, and then by two lines denoting a single standard deviation ($E[L_i] \pm h^2\sqrt{V[L_i]}$) under the assumption of L_i from the same distribution. h is the number of outage hours. The blue dashed Rayleigh lines are identical in all sub-figures, and are provided for reference.

The expected number of outages is given in the title of each sub-plot – the actual number of outages is of course different between each simulation round. The three red lines describe the results. Again, the one in the middle is the mean, whereas the two others represent $E[\sum L_i] \pm \sqrt{V[\sum L_i]}$, based on the empirical standard deviation of the simulated sample.

From the series of sub-figures a few observations can be made. First, the Rayleigh approximation seems to be adequate – its mean and standard deviation matches the simulations based on Dataset 1 for the situations with reasonably many outages. Second, as the number of outages decreases, the standard deviation grows – it resides somewhere in the interval between the limiting cases of (i) independent or (ii) identical approximations of Eq. 11. Third, when the expected number of outages is just a single one, the mean drops significantly. The reason is that a fair share ($e^{-1} \approx 37\%$) of the simulations will now have zero outages, with a corresponding zero cost of downtime. As seen in the graph, this change also entails a greater standard deviation.

B. Dataset 2

Dataset 2 represents trading volumes on the *Elspot* market of Nord Pool Spot – the largest physical power market in the world. The dataset was created by multiplying hourly energy volumes [MWh] with hourly prices [€/MWh] for 2009, both of which are readily available at the Nord Pool website (`www.nordpoolspot.com`). Figure 5 illustrates the dataset – a rare example of a business process running continuously, 24 hours day, 7 days a week, every week of the year. Other stock markets typically close and open every day. It comprises the hourly trading volume for every hour of 2009, except for four anomalous hours in the statistics provided – 8756 hours

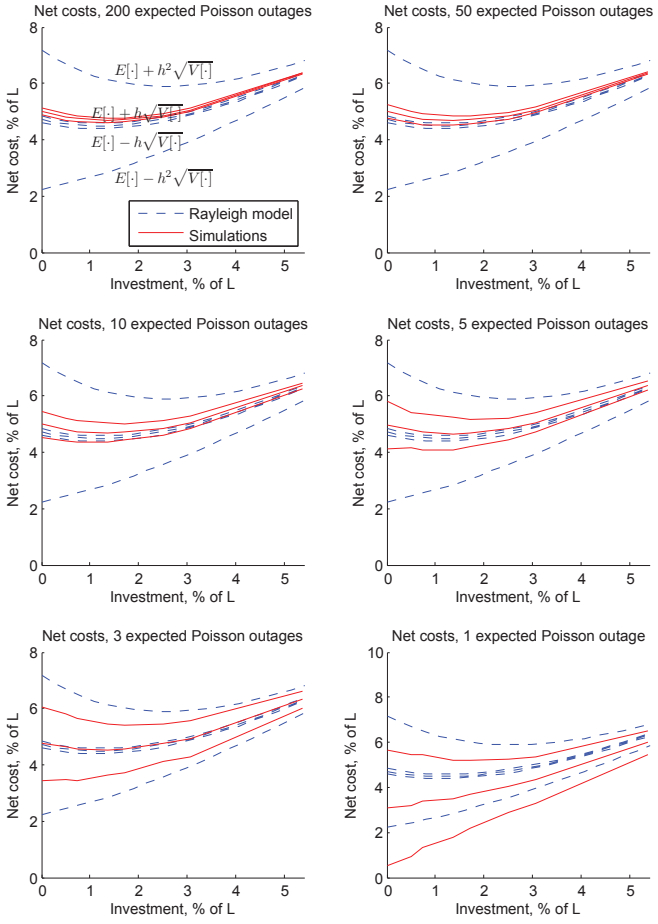


Fig. 6. Means and standard deviations as predicted by the Rayleigh model and as simulated from Dataset 2.

in all. As evident from Fig. 5, there is a seasonal variation – the trading volume is greater in the winter than in the summer.

A disadvantage is that the connection between the monetary trading volume and the actual downtime cost (as stipulated in Eq. 2) is somewhat disputable. However, even though a large trading volume does not necessarily mean a large downtime cost, there is some correlation. Even though the best measure of downtime cost on the power market might not even be monetary, the monetary volume hints at the amount of adjustments that are necessary to meet the balance requirements in the power grid. Furthermore, for the purpose of the article, we might disregard the context and just consider an interesting IT system, running 24/7, facilitating the transfer of large amounts of money (285.5 TWh were traded on the *Elspot* market in 2009, at an average price of 35.02 €/MWh), with readily available data.

To model L_i, we again select the Rayleigh distribution using quantile-quantile plots. For Dataset 2, the maximum likelihood estimate of the Rayleigh parameter gives $\sigma \approx 0.0086$. Figure 6 displays simulations for Dataset 2 corresponding to those for Dataset 1 in Fig. 4. Again, Monte Carlo simulations were made where each given availability investment level was simulated under conditions of different expected numbers of outages. Each of the cases was simulated 1 000 times.

Considering the two sets of simulations, it seems that the simple model for optimal availability investments epitomized by Eq. 5 is relatively robust in terms of means. The Rayleigh distribution approximation also appears adequate to yield more or less optimal investment levels c^*. This is true to the extent that the minima of modeled and simulated net costs coincide in Figs. 4 and 6, as they do rather nicely in most cases.

However, with a decreasing number of expected outages, the standard deviation grows. As observed in section IV, a prudent decision-maker would like to make an investment with a certain confidence interval, as expressed by the Chebyshev inequality (Eq. 9). Since both sets of simulations show that the standard deviation of the costs grows steadily as the number of outages decreases, the prudent decision maker might trade a somewhat higher expected cost for a somewhat lower standard deviation, i.e. prefer a solution with many short outages to a solution with fewer and longer outages, even if the first on average has slightly more downtime. Exploring the implications of this for optimal decision-making is the subject of the next section.

VI. Availability Management Strategies

A. Prudent decision-making

The simulations based on empirical data confirm that the decision problem of section III is a bit naive. To make optimal decisions, we need to take account both of availability A (as defined in Eq. 1) and the number of outages N (which might be modeled as in Eq. 12).

In section III, we implicitly took for granted that the relevant preferences concerning availability of IT systems are simply the expected monetary net cost, which is to be minimized. This makes perfect sense in the deterministic case, but it is not necessarily as sensible in the presence of uncertain outcomes, as our variance studies showed.

Figure 7 uses the cost distribution of dataset 2 to show simulated net cost outcomes with 95.5% availability (an arbitrarily chosen figure, non-optimal by Eq. 5) and 96.7% availability (the optimal figure by Eq. 5) for different expected numbers of Poisson outages. Each boxplot is based on 1 000 simulations. The relation between investment c and availability is that of function f_1 of Fig. 2, with $\alpha = 0.3$. Studying these figures, we can observe the same trends that were described in the previous section. As the number of outages drops, the *mean* expected net cost holds relatively steady, whereas the variance grows. For the cases of 5 or 3 expected outages, zero downtime simulations start to appear, lowering the *smallest costs* down to the baseline cost for having the expected average availability of 95.5% or 96.7%, respectively. On the other hand, the *largest costs* observed climb correspondingly. So, what should be preferred to what?

Decision rule 1 (Minimize expected net cost): Choosing between alternatives, the alternative with the lowest expected net cost should be selected.

Minimizing the expected net costs is popular and straightforward. With this strategy, variance does not matter, and the alternatives depicted in Fig. 7 are ordered from worst to best ($A \prec B$ means that B is preferred to A) like this:

$$(95.5\%, 10) \prec (95.5\%, 200) \prec (95.5\%, 50) \prec (95.5\%, 5) \prec (96.7\%, 10) \prec (95.5\%, 3) \prec (96.7\%, 200) \prec (96.7\%, 50) \prec (96.7\%, 5) \prec (96.7\%, 3)$$

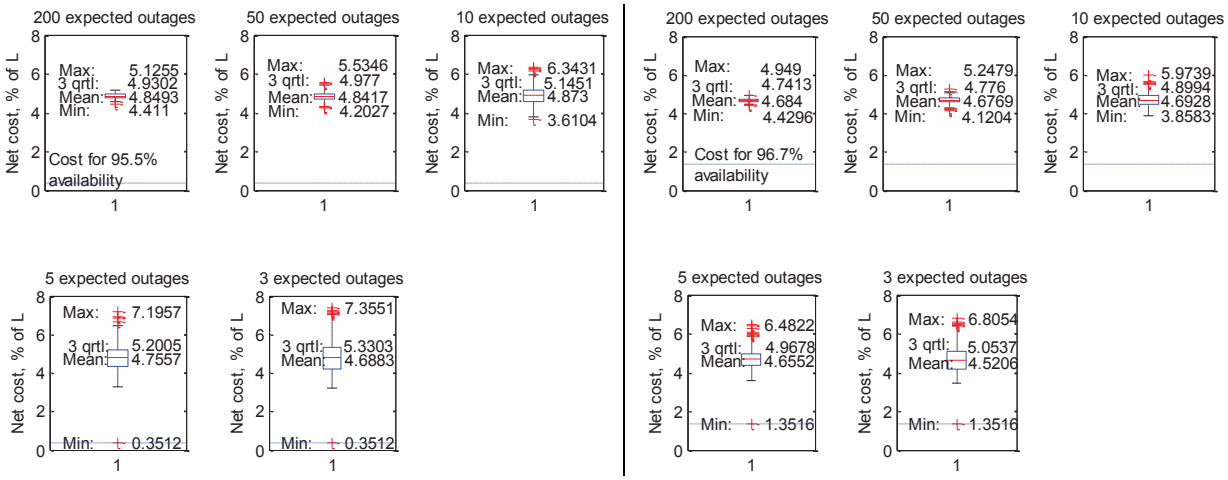


Fig. 7. Simulated net cost outcomes with 95.5% availability (left, non-optimal by Eq. 5) and 96.6667% availability (right, optimal by Eq. 5) for different expected numbers of Poisson outages with costs from dataset 2. Each boxplot is based on 1 000 simulations.

Of course, not all of the differences are statistically significant, but the principle is clear. On average, the expected cost of the "suboptimal" 95.5% availability alternatives is slightly higher than that of the "optimal" 96.7% availability alternatives – as is to be expected. However, this is not a strict rule: With only 3 outages, 95.5% actually perfomes better than 96.7% with 10 outages. Of course, another 1 000 simulations might give a different result, but this again underlines the importance of variance.

Minimizing expected cost derives much of its force from the argument that if it is adhered to in a long series of repeated cases, the result will be better than by any other strategy. However, when the number of outages approaches just a few, this argument loses some of its force. What good is it to be better off on *on average* if there are just a few outcomes, and those were bad? This is the rationale for more cautious decision rules:

Decision rule 2 (Minimize maximum net cost): Choosing between alternatives, the worst outcomes of all the alternatives should be evaluated, and the alternative with the best worst outcome, i.e. the smallest maximum cost, should be selected.

This rule, Minimax for short, is usually applied for decisions under ignorance, i.e. when no probabilities are known [39]. Nevertheless, it can be applied to the alternatives depicted in Fig. 7, ordering them from worst to best:

$$(95.5\%, 3) \prec (95.5\%, 5) \prec (96.7\%, 3) \prec (96.7\%, 5) \prec (95.5\%, 10) \prec (96.7\%, 10) \prec (95.5\%, 50) \prec (96.7\%, 50) \prec (95.5\%, 200) \prec (96.7\%, 200)$$

It is striking that the most preferred alternative under the Minimize expected cost rule, (96.7%,3), turns out to be the third worst alternative under the Minimax rule. With just three expected outages, they might occur at very critical times, and last for a long time (in this example, based on dataset 2, that corresponds to long outages during a cold snap, when electricity prices are very high). So, while favorable on average, (96.7%,3) also offers a large potential setback. Splitting the outage duration into 200 different occurences gives a much smaller variance.

Of course, Decision rule 2 might seem overly cautious and pessimistic: it decides based on the worst possible outcome. The worst outcome (i.e. the smallest maximum cost) is typically an outlier (the boxplots mark as outliers, with a red + symbol, any value situated more than 1.5 interquartile range above the third quartile or below the first). To accommodate for this, less outlier-sensitive cautious rules could be devised, where it is not the maximum costs that should be considered, but for instance the third quartile. Such rules will still be cautious, but not quite as pessimistic as the minimax rule.

The different decision rules and orderings again underline the fact that a single availability figure (e.g. 99%) is ambiguous and needs to be supplemented with the number of outages. The discussion so far indicates that among strategies that are equivalent in terms of availability as a percentage, it is prudent to select the strategy with many short outages rather than a few long. However, such a recommendation is only valid for particular scenarios, since the decision also highly depends on the type of business and service run. We now turn to a more formal consideration of this question.

B. Non-linear hourly outage costs

In section III, we make the simplifying assumption that the financial losses L incurred by an outage are simply proportional to the outage duration t. While useful as a first approximation, this is clearly not true. It is instructive to consider two archetypal cases of how this assumption fails:

Fixed restart cost For some services, each outage entails a fixed cost for restarting equipment and getting the business process back to where it was. The archetypal case is an IT service connected to a physical industrial process, e.g. a rolling mill. It is easy to understand that working temperatures in the 1 000 °C range and daily supply chains involving thousands of metric tons make each and every unplanned restart of sch a system very expensive.

Snowball effect For some services, short outages can go unnoticed. In a workshop conducted in May 2011, a number of practitioners from the enterprises

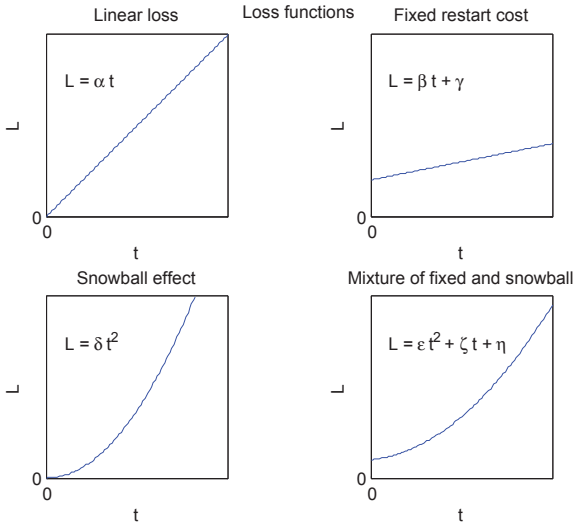


Fig. 8. Different conceptual loss functions L of outage duration t.

participating in the study described in [19] had the opportunity to discuss issues related to availability modeling and prediction. Consider three example given during the workshop: (i) An automated teller machine (ATM) out of bills to dispense can still deliberately be kept open for transactions. Customers will notice a closed ATM as being down, whereas the customer who cannot make a withdrawal from an open ATM will not necessarily consider it an outage. (ii) In Internet banking, a small hiccup will just make the customer reload the website, no harm done. (iii) A credit card transaction that fails will most often just result in the customer trying again. However, if the outage persists, not only will that particular transaction fail, but other customers standing in line behind will leave the store, (or stop using that ATM brand, or switch to a bank offering more reliable Internet banking) creating a snowball effect in which the losses by far surpass the loss of expected revenue during the outage itself.

Fig. 8 depicts different conceptual loss functions: First we have the linear assumption of losses proportional to outage duration, followed by the two archetypes, and to the lower right a mixture of the archetypes. The figure also suggests simple mathematical models for each case.

Clearly, the archetypal cases pull in different directions when it comes to the trade-off between shorter outages or fewer. With a fixed restart cost, it is preferable to have fewer outages, so that fewer restart costs will have to be paid. With a snowball effect, it is preferable to have shorter outages, so as to avoid the accelerating cost of longer outages. Again, this assumes a fixed level of availability. Using the mixed model of Fig. 8, a cumulative annual loss function is easy to find:

$$L = n\left(\varepsilon\left(\frac{h}{n}\right)^2 + \zeta\left(\frac{h}{n}\right) + \eta\right) \tag{13}$$

As before, h is the number of outage hours (e.g. 87.6 for 99% availability on a 24-7 service) and n is the number of outages these hours are distributed over. The optimal number of outages n^* is given by setting the derivative with respect to n to zero, i.e. the first order condition:

$$L_n = -\varepsilon\left(\frac{h}{n^*}\right)^2 + \eta = 0 \Rightarrow n^* = h\sqrt{\frac{\varepsilon}{\eta}} \tag{14}$$

n^* thus optimally balances the conflicting driving forces of the archetypal cases.

In a deterministic world, n^* would be the end of the story. However, we know from Section V that the ζ parameter – the outage cost per hour – exhibits a lot of variance. Re-using the Rayleigh approximation, we can again consider the two limiting cases of variance from Eq. 11. Fig. 9 depicts the cumulative annual loss function with three different parameter settings $(\varepsilon, \zeta, \eta)$ along with standard deviations plotted as in Figs. 4 and 6. Note that the inner limiting case almost coincides with the mean in the two first cases; only in the last case is it clearly visible.

In the leftmost figure, the snowball effect is comparably large to the fixed restart cost. With a low fixed restart cost, it might be prudent to aim for a *greater* number of outages than n^*, in order to make sure that the standard deviation is closer to the inner than the outer limiting case. This can be contrasted with the case depicted in the middle. Here, the fixed restart cost is much larger, so moving to more outages in order to reduce variance does not seem like a good idea. The rightmost figure again exhibits a sizable fixed restart cost. Here, however, the standard deviation in its outer limiting case is large enough to warrant a move to a number of outages greater than n^*, in order to keep *variance down*, even though this increases the *expected* total loss.

Thus, even with a known analytical model for the optimal number of outages, it might be worth to aim for a number of outages *greater* than n^*, considering variance, as is clearly shown in Fig. 9.

This particular analytical model should, of course, be taken with a grain of salt. However, the qualitative phenomena of *fixed restart cost* and *snowball effect* are real. This section shows that even in the presence of such effects, it is important to keep in mind the variance of hourly costs. Sometimes this variance overshadows non-linear effects, sometimes it does not.

C. Increase MTTF or decrease MTTR?

Eq. 5 gives a first order condition for an optimal investment in availability. To realize higher availability, such an investment can employ either one, or both, of two different strategies inferred from Eq. 1:

Increase the MTTF This will result in *fewer* outages per time unit, leading to higher availability. (A mechanism to achieve this might be the introduction of redundant system parts with automatic failover switching.)

Decrease the MTTR This will result in *shorter* outages whenever they occur, leading to higher availability. (A mechanism to achieve this might be to train repair crews better.)

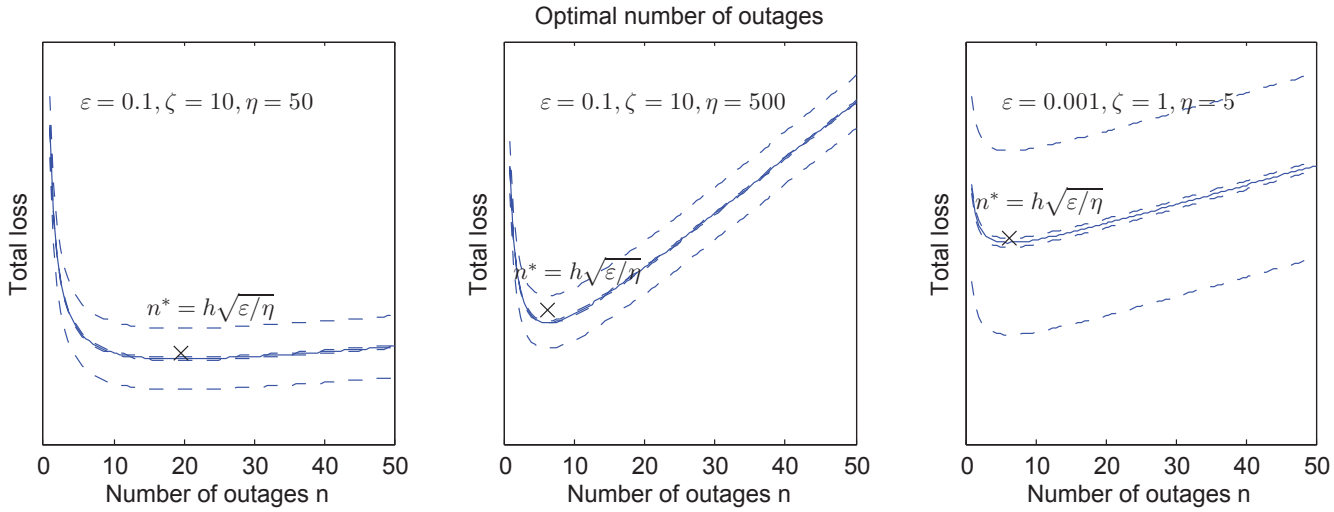


Fig. 9. Optimal number of outages n^* for a mixed loss function with different parameters. Standard deviations based on the Rayleigh approximation for ζ and on the two limiting cases of Eq. 11.

In the light of this, the discussion in this section can be summarized like this:

1) In the case of a *linear loss function*, i.e. the cost per downtime hour not being affected by the outage length, decreasing MTTR is preferable to increasing MTTF. The reason is the lower variance of the resulting availability and hence the lower worst-case cost. This is the basic case considered using empirical data and simulations in Section V.
2) In the case of a *fixed restart cost* loss function, increasing MTTF is generally preferable to decreasing MTTR.
3) In the case of a loss function with *snowball effect*, decreasing MTTR is generally preferable to increasing MTTF.
4) In the general case of a loss function with *both* fixed restart cost and snowball effect, Eq. 14 gives a first order condition for an expected optimal number of outages. However, the variance of the hourly outage cost can still make it prudent to aim for a greater number of outages than this expected optimum.

VII. Analysis

A. Feasibility

In order to make the approach described in the previous few sections as valuable as possible to decision makers, a number of functions, distributions and parameters must be obtained.

First, functions like the examples in Section 4 must be found. This task basically amounts to finding how much availability is increased when various actions are taken to improve it. While this is a difficult task, it is not impossible. In [18], an expert-based causal model is presented, where the relative importance of 16 factors affecting availability is accounted for. Using this model, it is possible to predict, quantitatively, to what extent a given action is likely to increase availability of an enterprise IT system. This framework is further detailed and tested in 9 industrial case studies in [19], where the model is also cast into a form based on standardized ITIL terminology. Renowned consultancy firm Gartner offers another model, quantifying the relation between investment and availability [40]. Here, the authors define a baseline cost X corresponding to a "standard IT service". Making such a service "highly available" (defined as a minimum of 99.3% availability) is assessed to cost $2.15X$. Moving up to "continuous availability" (99.81%) is assessed to cost $6.45X$. The final step on the Gartner ladder is dubbed "multisite continuous availability", deemed to cost $8.6X$. A third example of an investment-availabilty functional model is found in [24]. Even though this model makes assumptions somewhat different from requirements R1-R4, it nevertheless demonstrates the feasibility of finding the desired functional forms. To conclude, it is clear that functions such as 4 can indeed be found.

Second, downtime durations and distributions should be found. There are numerous papers that have empirically explored the repair rate of various types of failures, e.g. [41]–[43]. However, these contributions typically do not deal with enterprise level services, but rather with particular applications such as parallel computing, or with de-bugging software. However, in a recent study, restore times of enterprise IT services in a large Nordic bank were studied, based on a dataset of more than 2 000 incidents [44]. A lognormal distribution was found to best represent the distribution of restore times. It is instructive to consider a few examples: For a service with 848 downtime incidents to analyze, the 95% confidence intervals for the parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ of the lognormal distribution were $\hat{\mu} \in [4.41, 4.62]$, $\hat{\sigma} \in [1.50, 1.65]$. For another service, with just 49 downtime incidents in the sample, the corresponding estimates were $\hat{\mu} \in [3.44, 4.20]$, $\hat{\sigma} \in [1.11, 1.66]$. To conclude, it is clear that downtime durations and distributions can indeed be estimated with good precision, even on the enterprise service level.

Third, the fixed restart and snowball effects need to be quantified. Fixed restart costs are relatively transparent to the companies affected. Snowball effects, on the other hand, are typically more dependent on customer behavior, and thus harder to predict. A qualitative discussion of how service

managers can control customer perception of a queuing experience is found in [45], whereas [46] offers quantitative results on how reliability, among other factors, affects customer perceptions of e-service quality in online shopping. Another detailed analysis of customer behavior in face of delays is found in [47]. Using such sources, an estimate of the snowball effect can be obtained.

Fourth, an enlightened trade-off between MTTF and MTTR, as described at the end of Section VI, requires robust estimates of these figures. An interview-based method for obtaining such estimates is proposed and empirically tested in [48], where enterprise architecture models are used to estimate the availability of 5 enterprise services from the banking and electric utility industries. The annual downtime estimates found using this method were always within eight hours from the actual downtimes. Since no case study required more than twenty man-hours of work (including making the architectural models, conducting interviews, and make calculations), this demonstrates that accurate MTTF and MTTR values for enterprise IT services can be found relatively easy. This method is particularly useful if logs (cf. above) are not available.

B. Practitioner relevance and SLA management

Is it really the case that practitioners do not know how to invest to improve availability? Recall the workshop referred to above. The participants, mostly from the financial industry, agreed that even when action lists for improving availability are decided upon, no predictions are made of the consequences of individual actions. Clearly, this is an area where the results of this paper can give some guidance.

Furthermore, the workshop participants noted that many companies are immature when it comes to writing service level agreements (SLAs). This is critical, as the use of external services is becoming ever more common and their complexity is increased by having services delivered by sub-contractors to sub-contractors, i.e. service layering. The workshop participants noted that a lot of mistakes when it comes to SLA management are made not by the service provider, but rather internally. Poor requirements analysis can lead to fulfilled SLAs but dissatisfied users or customers.

One key lesson of the present paper is the importance of viewing availability as a more complex concept than a mere percentage. The number and duration of outages also need to be taken into account. However, when it comes to writing SLAs, well-regarded practitioner sources such as ITIL and Gartner offer a mixed bag, sometimes reflecting the full complexity of availability, sometimes not.

The importance of outage duration in SLAs is consistent with some Gartner recommendations such as [40], where the criticality of mean-time-to-restore-service (MTRS) is stressed, and [12] which includes response times (a key part of MTTR) as a key SLA element. However, in a Gartner report on cloud-based contact center services, while stressing SLA importance, only availability percentages are discussed [49]. Similarly, another report titled ”Best Practices for Service-Level Agreements for Software as a Service” fails to address the importance of outage duration [50].

The ITIL framework, in appendix F to the volume on service design, contains a sample SLA [34]. This SLA suggests that an availability target be specified as a percentage, and also that a maximum number of service breaks be stipulated. However, a maximum outage duration (i.e. a maximum MTTR) is not specified in the template.

As for the state of the practice, several trends can be observed. One finding very much in line with the verdict of the practitioner workshop is a Gartner survey where only 46% of the respondents in a 2008 questionnaire at all had formal availability SLAs for their mission-critical IT services [2]. On the other hand, the same survey showed that more than 50% of the audience had a Recovery Time Objective (RTO) of one hour or less for top-tier, mission-critical applications.

The main implications of our investigation for SLA management are the bullet points at the end of Section VI. However, when applying them, one should always be aware that there is a potential conflict of interest between the service provider and the service consumer. For example, even if the service consumer would like to achieve an availability goal with shorter outages, rather than fewer (following the arguments above), the service provider might find it more convenient to do the opposite. If the SLA is written only with respect to an availability percentage, a service provider might install new components with a better MTTF and simultaneously cut back on the staff doing restores. Thus the availability percentage is preserved, but not in the way most beneficial to the consumer.

Another important insight from Section VI is that no service level requirement should be taken for granted. An ambitious service level requirement from the business side should sometimes be negotiated down – it might not be worth the cost. This is well known. However, a mandatory service level requirement, such as one mandated by law, should sometimes be negotiated up – it might be worth the extra cost. *All* service level requirements ought thus to be scrutinized using Eqs. 5 and 14, without exception.

VIII. Conclusions

We have shown the importance of seeing availability not as a mere percentage, but as a more complex phenomenon, where the number of outages and their duration, as well as the variance of hourly downtime costs, play important roles.

The importance of variance was demonstrated using simulations on existing data sets of revenue data. An important implication is that when outage costs are proportional to outage duration, more but shorter outages should be preferred to fewer but longer, in order to minimize variance. Furthermore, we considered two archetypal cases where the cost of an outage depends non-linearly on its duration: (i) the case of fixed restart costs and (ii) the case of a snowball effect. An optimal outage length was derived for this case, but some necessary guidance was also given for how to apply this rule when the variance of hourly downtime costs is considered.

The mathematical nature of the analysis makes it suitable for implementation in a tool. Some feasibility issues were discussed in the paper, including a survey of how the necessary data could be acquired. While some information, such as functional forms, are suitable to be derived from academic publications, most of the data needs to be supplied as a business intelligence-solution, where statistics both from the

financial and the operations department are continuously fed and integrated for analysis. Implementing such a tool, and evaluating it in a corporate setting with an approprite design science methodology, is a possibility for future work. Judging from practitioner experience, one important area of application is writing service level agreements. The state of the practice is mixed, and far from all SLAs written consider availability to be more than a mere percentage.

Two additional important themes for future work have been identified: First, there is a need for more research on the functional relationship between investment in different solutions (independent variable) and resulting average availability and number of outages (dependent variables). The existing academic literature is quite void in this respect, but our results indicate that this is a crucial connection to investigate further.

Second, as outsourcing and service-orientation grow ever more popular, the relationship between service-providers and service-consumers ought to be investigated using game theory in the spirit of [51]. By acknowledging that service providers and service consumers can have conflicting as well as mutual interests, a game-theoretic approach can bring a lot of clarity to the activity of writing SLAs.

ACKNOWLEDGMENT

The author wishes to thank Johan König, who acted as the discussant of the first version of this paper at a doctoral research seminar at KTH, and Göran Ericsson, who made useful remarks on a later version. Furthermore, the three anonymous reviewers offered a number of comments that have substantially improved the quality of the paper.

REFERENCES

[1] IBM Global Services, "Improving systems availability," IBM Global Services, Tech. Rep., 1998.

[2] D. Scott, "How to assess your IT service availability levels," Gartner, Inc., Tech. Rep., Apr. 2009.

[3] U. Franke, R. Lagerström, M. Ekstedt, J. Saat, and R. Winter, "Trends in enterprise architecture practice—a survey," in *Proc. 2010 Trends Enterprise Architecture Research Workshop*.

[4] K. Harris, "Gartner CEO and senior business executive survey, 2010: perceptions of IT and tactical fixes," Gartner, Inc., Tech. Rep., Mar. 2010.

[5] A. Bharadwaj, M. Keil, and M. Mähring, "Effects of information technology failures on the market value of firms," *J. Strategic Inf. Syst.*, vol. 18, no. 2, pp. 66–79, 2009.

[6] S. I. Baek, S. J. Lee, and G. G. Lim, "Exploring impacts of IT service failure on firm's market value," in *Proc. 2008 International Conf. Netw. Comput. Advanced Inf. Manag.*, vol. 2, pp. 450–454.

[7] K. Campbell, L. A. Gordon, M. P. Loeb, and L. Zhou, "The economic cost of publicly announced information security breaches: empirical evidence from the stock market," *J. Comput. Secur.*, vol. 11, no. 3, pp. 431–448, 2003.

[8] J. Gray and D. Siewiorek, "High-availability computer systems," *Comput.*, vol. 24, no. 9, pp. 39–48, Sep. 1991.

[9] B. Butler and P. Gray, "Reliability, mindfulness, and information systems," *MIS Quarterly*, vol. 30, no. 2, pp. 211–224, 2006.

[10] N. Bajgoric, "Information technologies for business continuity: an implementation framework," *Inf. Manag. Comput. Security*, vol. 14, no. 5, pp. 450–466, 2006.

[11] M. Harper, C. Lawler, and M. Thornton, "IT application downtime, executive visibility and disaster tolerant computing," in *2005 International Conf. Cybernetics Inf. Technol., Syst. Appl.*

[12] D. Williams, "The challenges and approaches of establishing IT infrastructure monitoring SLAs in IT operations," Gartner, Inc., Tech. Rep., Sep. 2010.

[13] E. Marcus and H. Stern, *Blueprints for High Availability*, 2nd edition. John Wiley & Sons, Inc., 2003.

[14] F. Piedad and M. Hawkins, *High Availability: Design, Techniques, and Processes*. Prentice Hall, 2001.

[15] M. Malek, B. Milic, and N. Milanovic, "Analytical availability assessment of IT services," in *Proc. 2008 ISAS*, pp. 207–224.

[16] N. Milanovic, B. Milic, and M. Malek, "Modeling business process availability," in *Proc. 2008 IEEE Congress Services - Part I*, pp. 315–321.

[17] X. Zhang and H. Pham, "An analysis of factors affecting software reliability," *J. Syst. Softw.*, vol. 50, no. 1, pp. 43–56, 2000.

[18] U. Franke, P. Johnson, J. König, and L. Marcks von Würtemberg, "Availability of enterprise IT systems—an expert-based Bayesian framework," *Softw. Quality J.*, 2011, DOI: 10.1007/s11219-011-9141-z.

[19] U. Franke, P. Johnson, and J. König, "Availability of enterprise IT systems—empirical validation of a Bayesian framework," 2011, submitted manuscript.

[20] D. L. Weisman, "The emerging market for 'faultless' telecommunications," *Telecommun. Policy*, vol. 14, no. 4, pp. 333–341, 1990. Available: http://www.sciencedirect.com/science/article/B6VCC-4697WYV-1X/2/50a3ce44f77e4855cd4b4f0e00851802

[21] H. K. Cheng, "Optimal internal pricing and backup capacity of computer systems subject to breakdowns," *Decision Support Syst.*, vol. 19, no. 2, pp. 93–108, 1997, Economics of Inf. Syst. Available: http://www.sciencedirect.com/science/article/B6V8S-3SWV8BD-3/2/89724005a4721760e8465dd717ff69d0

[22] ——, "Pricing and capacity decisions of clustered twin-computer systems subject to breakdowns," *Decision Support Syst.*, vol. 25, no. 1, pp. 19–37, 1999. Available: http://www.sciencedirect.com/science/article/B6V8S-3VY097F-2/2/07d393ae4f0bf2d700ce14ae2fee2857

[23] N. Carr, "IT doesn't matter," *IEEE Eng. Manag. Rev.*, vol. 32, no. 1, pp. 24–32, 2004.

[24] M. Wiboonrat, "An optimal data center availability and investment trade-offs," in *Proc. 2008 International Conf. Softw. Eng., Artificial Intelligence, Netw., Parallel/Distributed Comput.*, pp. 712–719.

[25] A. P. Sanghvi, "Optimal electricity supply reliability using customer shortage costs," *Energy Economics*, vol. 5, no. 2, pp. 129–136, 1983. Available: http://www.sciencedirect.com/science/article/B6V7G-458XJWV-7P/2/9be7230bf645a3bae9b59a12df967b27

[26] S. Burns and G. Gross, "Value of service reliability," *Power Syst., IEEE Trans.*, vol. 5, no. 3, pp. 825–834, Aug. 1990.

[27] A. Tishler, "Optimal production with uncertain interruptions in the supply of electricity: estimation of electricity outage costs," *European Economic Rev.*, vol. 37, no. 6, pp. 1259–1274, 1993. Available: http://www.sciencedirect.com/science/article/B6V64-45CX08M-5Y/2/0a8ccec4ea60a84fa0b872452417497f

[28] F. Carlsson and P. Martinsson, "Does it matter when a power outage occurs? A choice experiment study on the willingness to pay to avoid power outages," *Energy Economics*, vol. 30, no. 3, pp. 1232–1245, 2008. Available: http://www.sciencedirect.com/science/article/B6V7G-4NN1TMP-2/2/75b0a0d26750099f78e3dc3e35a1af17

[29] M. de Nooij, C. Koopmans, and C. Bijvoet, "The value of supply security: the costs of power interruptions: economic input for damage reduction and investment in networks," *Energy Economics*, vol. 29, no. 2, pp. 277–295, 2007. Available: http://www.sciencedirect.com/science/article/B6V7G-4KBX4M9-2/2/b6df99655f2584c909982e3333c663e6

[30] T. Setzer, K. Bhattacharya, and H. Ludwig, "Change scheduling based on business impact analysis of change-related risk," *IEEE Trans. Netw. Service Manag.*, vol. 7, no. 1, pp. 58-71, 2010.

[31] C. Courcoubetis and V. Siris, "Managing and pricing service level agreements for differentiated services," in *1999 International Workshop Quality Service.*

[32] J. J. Trienekens, J. J. Bouman, and M. van der Zwan, "Specification of service level agreements: problems, principles and practices," *Softw. Quality J.*, vol. 12, pp. 43–57, 2004, 10.1023/B:SQJO.0000013358.61395.96. Available: http://dx.doi.org/10.1023/B:SQJO.0000013358.61395.96

[33] A. Høyland and M. Rausand, *System Reliability Theory: Models and Statistical Methods*. Wiley, 1994.

[34] S. Taylor, V. Lloyd, and C. Rudd, *Service Design (ITIL)*, the Stationery Office, 2007.

[35] D. Patterson, "A simple way to estimate the cost of downtime," in *Proc. 2002 Systems Administration Conf.*, pp. 185–8.

[36] L. Gordon and M. Loeb, "The economics of information security investment," *ACM Trans. Inf. Syst. Security*, vol. 5, no. 4, pp. 438–457, 2002.

[37] M. Rausand and A. Høyland, *System Reliability Theory: Models, Statistical Methods, and Applications*, 2nd edition. John Wiley & Sons, Inc., 2004. Available: http://www.ntnu.no/ross/srt

[38] F. Bergström and J. Arnberg, "När handlar vi? - Om konsumtionsmönster i den traditionella detaljhandeln," Handelns Utredningsinstitut, Tech. Rep., Mar. 2005 in Swedish.

[39] M. D. Resnik, *Choices: An Introduction to Decision Theory*. University of Minnesota Press, 1987.

[40] B. Malik and D. Scott, "How to calculate the cost of continuously available IT services," Gartner, Inc., Tech. Rep., Nov. 2010.

[41] S. Gokhale and R. Mullen, "A multiplicative model of software defect repair times," *Empirical Softw. Eng.*, vol. 15, pp. 296–319, 2010, 10.1007/s10664-009-9115-y. Available: http://dx.doi.org/10.1007/s10664-009-9115-y

[42] M. Vineyard, K. Amoako-Gyampah, and J. Meredith, "Failure rate distributions for flexible manufacturing systems: an empirical study," *European J. Operational Research*, vol. 116, no. 1, pp. 139–155, 1999.

[43] S. Matz, L. Votta, and M. Malkawi, "Analysis of failure and recovery rates in a wireless telecommunications system," in *Proc. 2002 IEEE Conf.Dependable Syst. Netw.*, pp. 687–693.

[44] U. Franke, H. Holm, and J. König, "On the distribution of time to recovery of enterprise IT services," 2011, submitted manuscript.

[45] M. Davis and J. Heineke, "Understanding the roles of the customer and the operation for better queue management," *International J. Operations Production Manag.*, vol. 14, no. 5, pp. 21–34, 1994.

[46] G. Lee and H. Lin, "Customer perceptions of e-service quality in online shopping," *International J. Retail Distribution Manag.*, vol. 33, no. 2, pp. 161–176, 2005.

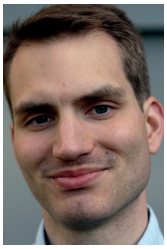
[47] J. Chen and N. Zhang, "Customer incentives in time-based environment," in *Service Enterprise Integration*, ser. Integrated Series Inf. Syst., R. Sharda, S. Voß, and C. Hsu, editors. Springer US, 2007, vol. 16, pp. 103–129, 10.1007/978-0-387-46364-3_4. Available: http://dx.doi.org/10.1007/978-0-387-46364-3_4

[48] P. Närman, U. Franke, J. König, M. Buschle, and M. Ekstedt, "Enterprise architecture availability analysis using fault trees and stakeholder interviews," *Enterprise Inf. Syst.*, 2011, accepted subject to revisions, to appear.

[49] D. O'Connell and D. Kraus, "Critical elements of cloud-based contact center services: pricing, service-level agreements and service integration," Gartner, Inc., Tech. Rep., June 2010.

[50] B. Pring, C. Ambrose, W. Maurer, and A. Bona, "Best practices for service-level agreements for software as a service," Gartner, Inc., Tech. Rep., Nov. 2010.

[51] H. Demirkan, M. Goul, and D. S. Soper, "Service level agreement negotiation: a theory-based exploratory study as a starting point for identifying negotiation support system requirements," *2005 Hawaii International Conf. Syst. Sciences*, vol. 1.



Ulrik Franke received his MSc in Engineering Physics from the Royal Institute of Technology (KTH) in Stockholm in 2007, and is now pursuing a Ph.D. in Industrial Information and Control Systems. His primary research interest is Enterprise Architecture (EA), a method for rational decision making in the realm of enterprise information systems, focusing on availability modeling of enterprise information systems.