Proc. Second International Symposium on Autonomous Decentralized Systems, Phoenix, AZ, April, 1995

Dependability Assessment for Decentralized Systems

Herbert Hecht Myron Hecht
SoHaR Incorporated
Beverly Hills, California

Abstract

Conventional dependability measures, such as reliability or availability, assume that the equipment characterized by the measure is either operational or has failed. This dichotomy does not hold for decentralized or distributed systems because these can operate in modes in which partial or degraded service is furnished. Whether a specific degraded mode should be counted as "operational" or "failed" is a subjective decision, but this decision can make a large difference in the dependability assessment. Examples show that this will affect not only numerical reporting, but also the selection of reliability improvements. A weighted averaging of the dependability measures obtained under various failure criteria is seen to be a workable method for reliability assessment that provides much more stable measures than can be obtained by selection of a single failure criterion.

1. Introduction

It is usually taken for granted that dependability is a desirable attribute for any computer based system, and particularly for distributed or decentralized systems. Indeed, the choice of a distributed system architecture is frequently motivated by dependability considerations. In practice it is difficult to quantify dependability for such systems because they can operate in multiple states, each with its own failure criterion. The purpose of this paper is to call attention to the need for further research in this area and to propose avenues for this research. Because of the large amount of resources devoted to achievement and verification of dependability, it is expected that there will be a large payoff from such efforts.

Dependability is defined as the trustworthiness of a system (in the present context a computer-based system), such that reliance can justifiably be placed on the service it delivers [1]. Dependability is commonly interpreted as incorporating four attributes: availability, reliability, safety, and security [2]. Not all of these are required in each application, but even with a focus on a single one of these attributes, assessment can be difficult.

The concepts and definitions for each of the dependability attributes were developed for systems or components that operate of fail, with a clear distinction between these states. Distributed and decentralized systems usually can operate in degraded states, with one or more elements failed, but sufficient others available to furnish some service. To represent the value of a system to the user, dependability measures therefore must account for the existence of states in which less than full capability is furnished.

2. The baseline system

An example of a typical decentralized system is shown in Figure 1. The system consists of a dispatch station and three identical highly autonomous service stations.

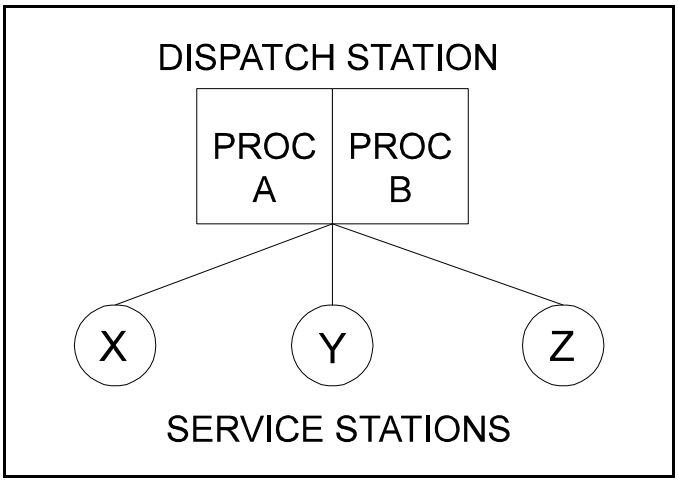

Figure 1 Baseline system

The latter may be automated manufacturing areas, military surveillance installations, or monitoring systems

for a nuclear power plant. Service requests (shop orders, target identifications, or instrument readings) arrive at the dispatch station in a random manner but with negligible probability of exceeding the allowable queuing time. The dispatch station routes the request to the first available server. In the baseline it is assumed that each service station has the capacity for providing 0.45 of the required output. Thus, with all stations operative, there is spare capacity, and if one station is down the system can still furnish 90 percent of the required output.

The service stations are dispersed and do not have resident maintenance personnel. Repair or replacement of a faulty component usually involves a downtime for that station of 24 hours. To simplify matters, dependability will in the following be expressed by a single attribute: availability (defined in [2] as "dependability with respect to readiness for usage"). For a given service station this is expressed as the probability that it is in an operative state. The baseline dispatch station consists of redundant processors (each with its own communication equipment) for which an availability of 0.98 per processor is specified. It is assumed that the availability of the processors is independent, and thus the overall availability of the dispatch station, A_d is 0.9996. The baseline availability of each service station, A_s is specified as 0.94.

The probability that the entire system with 3 service stations will be available is

$$A_3 = A_d \times A_s^3 = 0.830525 \qquad (1)$$

The capability for that mode is

$$C_3 = 1.0$$

Of course, the system can still perform at a nearly adequate level with only two service stations operative. For that mode

$$A_2 = 3 \times A_d \times A_s^2 \times (1 - A_s) = 0.15905$$

and the capability is

$$C_2 = 0.9.$$

The probability that the system can furnish *at least* this capability is

$$A_2' = A_3 + A_2 = 0.98924 \qquad (2)$$

Finally, we might be interested in the availability at the most degraded mode, in which only a single service station is operative, given by

$$A_1 = 3 \times A_d \times A_s \times (1 - A_s)^2 = 0.01015$$

for a capability of

$$C_1 = 0.45.$$

The probability that the system can furnish *at least* this capability is

$$A_1' = A_2' + A_1 = 0.99938 \qquad (3)$$

After having computed availability predictions ranging from 0. 83025 to 0.99938, the question arises: which is the true (the most meaningful, valid, or concrete) value of availability? This is an example of the difficulty encountered in dependability assessment of decentralized systems that was mentioned in the introduction, and this issue will now be explored. Some solutions will appear more appropriate than others, but it will be seen that the choice between the alternatives will always be subjective.

First assume that the service stations represent automated machine shops and that the capabilities, C_n, are measured by the number of widgets that can be produced per day. Since losing 10% of production for one day may be quite acceptable in this environment, the decision maker will probably adopt C_2 as his criterion and assess A_2' = 0.98924 as the most meaningful measure of dependability for the system.

Next, consider that the service stations represent air defense sensors, and that the capability is measured by the number of targets that can be tracked. Since military decision makers must contend with an adversary who might utilize intelligence about the status of the system to launch an attack when it is in a degraded mode, they will not agree that a 10% loss in capability is tolerable ("if we thought that would be sufficient, we would have specified it"). They will thus regard A_1 = 0.83025 as the most representative availability prediction and will probably assert that an improvement in availability is desirable.

For a safety system for a nuclear power plant, the service stations represent sensor processing of particle flux or thermodynamic quantities, and the capability is measured by the number of sensor samples that will be processed per unit time. Since the multiplicity of sensor readings is needed primarily to tolerate sensor outages, and since these failures can be assumed to occur randomly rather than at the volition of a malicious enemy, even the 45% processing capability corresponding to C_1 may be acceptable, and the decision maker may consider A_1' = 0.99938 as a representative availability (but probably not as sufficiently high as discussed below).

Table 1. Results of alternative availability improvements

Improvement Type	C_3	A_3	C_2	A_2'	C_1	A_1'
(a) Increased serv. sta. capability	1.0	.83025	1.0	.98924	0.5	.99938
(b) Increased serv. sta. availability	1.0	.91231	0.9	.99696	0.45	.99957
(c) Increased disp. sta. redundancy	1.0	.83058	0.9	.98962	0.45	.99978
Combination of (b) and (c)	1.0	.91267	0.9	.99735	0.45	.99997

It is seen that even for a specific instance of dependability, such as the availability metric used in this example, some subjective decisions enter into the assessment. The key issue is that a failure criterion must be specified. The following section shows that the availability assessment not only produces numbers that characterize a system, but that these numbers can have significant effects on system design and improvement decisions.

3. Availability improvement

Availability assessment (or dependability assessment in general) is used to support decision processes. One decision outcome is that the availability is adequate for the intended purpose, in which case the only further activity will be to monitor that it does not drop below the required value. The other outcome is a finding that the availability needs to be improved. The following paragraphs show how the selection of the most suitable improvement type is affected by the capability and availability choices made in each of the above cases.

Three possible modifications of the baseline system may be used to improve the availability:

(a) upgrade the capability of each service station from 0.45 to 0.5

(b) improve the availability of each service station, A_s, from 0.94 to 0.97

(c) add a third dispatch computer, making the dispatch station triple redundant.

For the time being it is assumed that each of these improvements has the same cost. Which one should be selected? To help answer this question, the capabilities and availabilities resulting from these improvements are shown in Table 1. They are computed by equations (1) - (3) of Section 2.

For the manufacturing application the failure criterion was the point at which production dropped below 90%, and for that criterion the single improvement that provides the highest availability is (b). The improvement in A_3 that is achieved thereby is also significant for this application.

For the air defense application the failure criterion is the point at which capability drops below 1.0 The highest availability for that case is achieved by improvement type (a). For the nuclear reactor safety system it is the point at which the capability drops below 45%, and therefore alternative (c) may be selected, although this provides practically no benefit under any of the other criteria. Because the availability for nuclear safety applications is usually required to be higher than 0.999, a very attractive improvement for this case is the combination of (b) and (c) shown in the last row of the table. This combination provides only very marginal improvement by the higher capability criteria but is decisively better at the C_1 level.

At this point we have seen that the selection of the failure criterion not only affects the reporting of a numerical value for availability but can also affect very important decisions for product improvement.

4. Integrated measures

In the above examples each of the decision makers adopted a different criterion, and it was acknowledged that this was a subjective act. It may be questioned whether the decisions can be improved by combining

Table 2. Selection of improvement by weighted measures

Improvement Type	W_1 W_2 W_3	.33 .33 .33	.2 .3 .5	.1 .3 .6	.5 .3 .2
(a) Increased serv. sta. capability		.93023	.91177	.89486	.96251
(b) Increased serv. sta. availability		.95992	.95516	.94643	.98134
(c) Increased disp. sta. redundancy		.93059	.91213	.89521	.96289
Combination of (b) and (c)		.96030	.95553	.94680	.98172

the criteria, possibly giving them different weights. An example of weighted measures is shown in Table 2, where W_n (n = 1..3) is the weight assigned to the availabilities associated with criterion C_n in Table 1 (W_1 is the weight assigned to the configuration in which only one node remains operative).

The first numerical column represents a "flat" weighting; the second one a "realistic" weighting (the system will spend most of its time with all three nodes operative, and thus assigning a higher weight to that state is justified); the third column represents an even heavier weighting of the fully operational state and might be termed "optimistic"; and the final column represents a "defensive" posture by over-weighting the operational state just short of failure. A significant finding is that the relative ranking of the improvement alternatives is independent of the assigned weights. Among the first three rows (b) is always highest, followed by (c) and (a). Also, the availability increase of the combination improvement over (b) alone evaluates to between 0.0003 and 0.0004 in all cases. It is thus seen that decisions (at least in this case) are independent of the weighting scheme over a fairly wide range, and also in that they show relative constancy in the assessment of incremental

Table 3. Selection of improvement by doubly weighted measures

Improvement Type	**W_1* W_2* W_3***	**.33 .33 .33**	**.2 .3 .5**	**.1 .3 .6**	**.5 .3 .2**
Original system		**.71619**	**.77216**	**.81022**	**.65801**
(a) Increased serv. sta. capability		**.76533**	**.81184**	**.84489**	**.71267**
(b) Increased serv. sta. availability		**.74560**	**.81530**	**.86155**	**.74560**
(c) Increased disp. sta. redundancy		**.71748**	**.77247**	**.81054**	**.71648**
Combination of (b) and (c)		**.74589**	**.81562**	**.86189**	**.74589**

Table 4. Weighting for availability increments

TABLE NO.	ΔA for improvement (b)			ΔA for improvement (c)		
	Highest	Lowest	Ratio	Highest	Lowest	Ratio
1	0.08206	0.00019	10.8	0.00040	0.00033	1.3
2	0.05157	0.01882	2.7	0.00038	0.00035	1.1
3	0.08759	0.02940	2.9	0.00038	0.00035	1.1

benefits.

The above approach to weighting suppresses one of the factors that led to the preference for (a) in the case of the air defense system, viz. the extension of 1.0 capability to the configuration with only two operable nodes. Double weighting, in which the second weight is the capability that exists at a given level, overcomes this problem. In Table 3 below, $W_n^* = W_n \times C_n$ (n = 1..3).

In Table 3 the first row contains the weighted availability for the original system. This information was identical to that for improvement alternative (a) in Tables 1 and 2 and was therefore not separately provided there.

The entries in this table do not represent availability by the classical definition because they are based on a product of capability and dependability. This product is sometimes referred to as *effectiveness* [3], and for some USAF applications it has been identified as the preferred measure for *system effectiveness* [4]. For the sake of simpler terminology we will include this measure under a broadened interpretation of availability and use the associated symbols because in the current context the distinction is not material. For flat assigned weights (first numerical column), the double weighting indicates an advantage for improvement type (a). This is in accordance with the expectations. It may at first appear surprising that this advantage prevails only with flat weighting, but this is explained by observing that in this scheme the weight assigned to criterion 2 is higher than in any other scheme, and that improvement type (a) provides benefits primarily under criterion 2. This explanation can be verified by setting $W_3 = W_2 = 0.4$ in the second numerical column, in which case the availability for (a) 0.82773 while it decreases for (b) to 0.81379.

The difference between the numerical entries in a given column is considerably less for the integrated measures than for the single criterion availability values shown in Table 1. This is to be expected since the integrated measures represent a melding of the individual values. The reduced differences become important when cost trade-offs are conducted, as shown in the following section.

5. Cost trade-offs

In the preceding section the cost of availability improvements was not explicitly considered, although in practice it is a major factor in the decision process. The key question that usually arises in this connection is "What cost is warranted for an improvement?" A simplified approach to this question for a conventional system (one that has only one operational state) is to determine the loss incurred due to downtime for the baseline system, and to multiply this by the reduction in downtime (or increase in availability) expected from the improvement under consideration:

$$V_i = \Delta A_i \times H \times L \qquad (4)$$

where V_i represents the value (the maximum allowable expenditure) of improvement *i*, ΔA_i represents the availability increment due to that improvement, H the expected number of operating hours over a defined period, and L the loss due to one hour of downtime. When this approach is tried in the decentralized environment difficulties arise because the loss as well as the availability increment depend on the capability level, and that does not expressly appear in the equation. When the above relation is applied to improvement (a) in Table 1, the result evaluates to zero because the change in capability is not reflected in ΔA. Another problem affecting all improvements is the large difference in ΔA, depending on which failure criterion or capability level is selected.

These difficulties are largely overcome by expressing ΔA in one of the weighted measures, particularly the

doubly weighted measure in which capability is directly factored in. Thus, in Table 3 all improvement alternatives show an increase in weighted availability over the original system under all weighting selections. Further, while the discrimination in availability between the rows (alternatives) remains reasonably constant, the variability of the improvement estimates (ΔA) between columns has been greatly reduced as shown in Table 4.

The "highest" and "lowest" entries were obtained by subtracting the first row from the row corresponding to the improvement (b or c) in each column of the previous tables, and then selecting the highest and lowest values of these differences. In all cases the ratio between the highest and the lowest in the first row is greater than that shown in the later rows. This indicates the ability of the integrated measures to furnish a less volatile indicator of the value of an improvement, and thus to furnish more useful information for cost trade-offs.

Difficulties are sometimes experienced in defining a value for the loss factor, L, in eq. (4). Among the three applications discussed earlier, the most direct determination of L will probably be found in the manufacturing environment. The major elements are unproductive labor, plant overhead, and loss of profit on the product, all of which are usually well known. But even in the other cases a rough estimate of the cost of downtime can be obtained. If the air defense system is down, a higher state of alert may be required for defensive squadrons (and the cost of that can be assessed), and if the plant safety system is down, the nuclear power plant will have to be shut down (or operate at reduced output for which an alternate monitoring system is adequate). Thus a creditable methodology is available for cost trade-offs for decentralized systems.

6. Summary and conclusions

Dependability assessment of decentralized systems must take into account that these systems can operate in a number of states for which significantly different dependability measures will be assessed by conventional methods. The paper has shown that integrated measures, based on weighted sums of the dependability for individual states, can be used to provide a single meaningful indicator of dependability, and in particular can guide the selection of improvement alternatives with or without cost considerations. Preferences between improvement alternatives were shown to be largely insensitive to the weights assigned to individual states.

In the examples presented here, availability was taken as a representative dependability characteristic, but others, such as reliability, security or safety can use the same methodology. The key conclusion is that there is a workable, though heuristic, approach to dependability assessment. It is recommended that experimentation with the methodology by undertaken by the research community, that it be refined, and ultimately used in the assessment of operational systems and for the selection of design or improvement alternatives.

References

[1] W. C. Carter, "A time for reflection", *Proceedings of the 12th IEEE Symposium on Fault Tolerant Computing (FTCS-12)*, p. 41, June 1982

[2] J. C. Laprie (ed.), *Dependability: Basic Concepts and Terminology,* Springer Verlag, 1992

[3] DOD, *Military Handbook, Electronic Reliability Design Handbook*, MIL-HDBK-338-1A, Volume 1, 1988

[4] Air Force Systems Command, *Chairman's Final Report, Weapon System Effectiveness Industry Advisor Committee (WSEIAC),* AFSC-TR-65-6 (AD-467816), 1965