Analysis and Justification of Indicators for Quality Assessment of Data Centers

Rosen Radkov

Abstract –**The quality of aDataCenter (DC) can be assessed by different criteria. The purpose of this article is to analyse the indicators that can be used to assess the quality of DC from the point of view of the accessibility of the services provided.Based on the analysis is substantiated set of significant indicators for the DC quality. Quantification of the quality is offered using a complex generalizedindicator, which can be used to benchmark between different DC's and selects an optimal variant for concrete business case.**

Keywords –**Data center, Quality, Dependability, Availability, Reliability.**

I. Introduction

New IT services like cloud computing are currently reaching a growing number of users. These new trendsare the reason for correspondingly high demands. Servers and data centers must be 100% available around the clock. Down times and interruptions or delays in data transmission must be virtually eliminated.

Which are the indicators for assessing the quality of a ***Data Center(DC)***? How to choose which DC'sIT infrastructure (ITIS) is right for concrete business case?

II. Terms and Definitions

A. Definition Quality

According to Wikipedia, the qualityis an essential attribute of any product or service and plays a crucial role in the function of the economic operators and market relations.

A modern definition of ***quality*** derives from Joseph Juran's "fitness for intended use." This means that quality is ***"meeting or exceeding customer expectations."***

Accordingly to that definition, the quality of a DC would be determined by the services provided to consumers and their criteria for satisfaction.

This article will not address issues related to the response time after to a request to the DC, its performance or energy efficiency.

B. Dependability

The quality of a DC can be assessed by different indicators. Commonly used indicator for this purpose in the technics is the concept of ***dependability***. In the most general sense the dependability is the ability of anobject to retain its essential properties in set modes and in operating conditions [1].

The systematic expositionof the dependability concept includes three parts: the threats to dependability, the attributes of dependability and tools for achieving dependability (Fig. 1) [2].

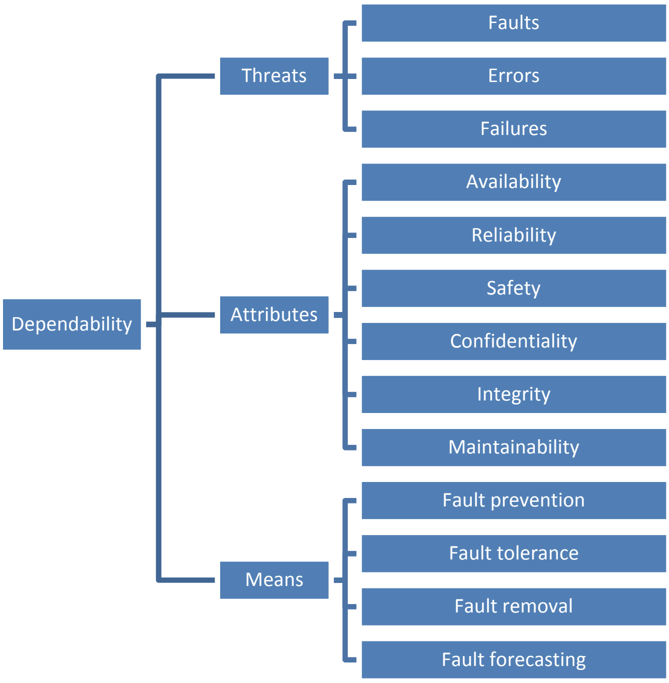

Fig. 1. The dependability tree

Dependability *attributes* describe the propertieswhich are required of a system. Dependability *threats* express the reasons for asystem to cease to perform its function or, in otherwords, the threats to dependability. Dependability *means* are the methods and techniques enabling the development of adependable system, such as fault prevention, fault tolerance, fault removal, and fault forecasting [3].

Dependability *threats* are defined as follow: a ***fault*** is physical defect, imperfection, or flaw that occurs in some software or hardware component; an ***error*** is a deviation from correctness or accuracy in computation, which occurs as a result of a fault; a***failure*** is a transition from correct to incorrect service, and this means that at least some external state of the service deviates from the correct state.

A common feature of the three terms is that they give us a message that somethingwent wrong. The difference is that, in the case of a fault, the problem occurred onthe physical level; in the case of an error, the problem occurred on the computationallevel; in the case of a failure, the problem occurred on a system level [3].

Rosen Radkov is a member of the Department of Software and Internet Technologies at Technical University of Varna, 1Studentska Str., Varna9010, Bulgaria, E-mail: rossen@actbg.bg

The DC is a set of interacting components where the dependability of the system is a complex property. The basic attributes of dependability are: ***availability***- readiness of the DC for correct service in any point in time [1], [2]; ***reliability*** - continuity of a correct service [1], [2]; ***safety*** - the absence of catastrophic consequences for the user(s) and the environment; ***confidentiality*** - protection from unauthorized disclosure; ***integrity*** - protection against unauthorized modification of information; ***maintainability*** - the ability of DC to operate without interruption during processes of repairs, maintenances and modifications;

If ***t*** is designated as atime interval in which a DC should be available, then ***reliability*** is the probability that DC will work according to its functional specification in period ***t***, and ***availability***is the probability that DC is not defective or under repair when it should be used. It can therefore be concluded that the availability is a function of the reliability and maintainability.

Dependability means are the methods and techniques enabling the development ofa dependable system.The means to attain dependability and security can be defined as follow:***fault prevention***: means to prevent the occurrence or introduction of faults; ***fault tolerance(FT)***: means to avoid service failures in the presence of faults; ***fault removal***: means to reduce the number and severity of faults; ***fault forecasting***: means to estimate the present number, the future incidence, and the likely consequences of faults.

C. MTBF and MTTR

These two terms, used in theory of dependability, have a place there. *Mean Time Between Failures (**MTBF**)* is defined as the average or expected time between two failures of a component or a system and *Mean Time To Recover (**MTTR**)* is the average time to recover a failed module or system. The following equations illustrates the relations of MTBF and MTTR with reliability and availability:

$$Reliability=e^{-\frac{t}{MTBF}} \tag{1}$$

$$Availability=\frac{MTBF}{MTBF+MTTR} \tag{2}$$

Analyzing these formulas we can draw the following conclusions [4]: First: if value of MTBF is higher, the higher the reliability and availability of the system; Second: MTTR affects availability. This means if it takes a long time to recover a system from afailure, the system is going to have a low availability; Third: High availability can be achieved if MTBF is very large compared to MTTR [5], [6], [7].

D. Disaster Tolerance and Disaster Recovery

DC and its ITIS need to become resistant to disasters.After a detailed analysis of the literature it is found that the characteristic „resistant to disaster DC"connects to the following two notions: ***disaster tolerance (DT)*** and ***disaster recovery (DR)***. The first concept is a characteristic of DC related to the disaster resilience, i.e. DC continues to carry out its activity, regardless of disaster. The second term is associated with the ability of DC to be recovered in minimal time and with minimal data loss disaster [5].

According to some sources that make analysis and design of ITIS, resistance to disaster (***DT***) isa system that in its characteristics corresponds to the DT. Other sources consider the recovery issues of ITIS after distress [5]. This leads to an incomplete assessment of the characteristics of ITIS, which are: current situation, the impact of the disaster on the different subsystems and vulnerabilities leading to potential risks.

*Therefore, to obtain more accurate and comprehensive description of the sustainability of ITIS disaster stages of planning, construction, research and evaluation, is necessary to use a comprehensive approach. It should consider both concepts - DT and DR, where both should be united in one concept for**Highly Reliable Data Centre (HRDC)**.*

E. RPO and RTO

A DC which retains its ability to provide services during disasters and to store the critical data post disaster will noted as ***Highly Reliable Data Centre (HRDC)***. In case of a fault on one / some of its components and inconsequence run unavailable, HRDC should be able to resume work for a minimum period of time. A solution that meets the requirements for HRDC is related to the value definition of two parameters: RTO (***Recovery Time Objective***) and RPO (***Recovery Point Objective***). The time needed to restore the work of a DC after a disaster (incident) is defined as RTO, while with RPO is being indicated the time interval before the disaster, which the business has accepted as eligible for which data will be lost (Fig.2).

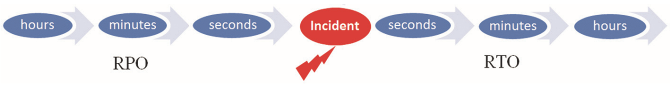

Fig. 2. Graphical representation of RPO and RTO

III. Indicators for Assessment of DCs

In the case of DC design, taking into account the requirements of the customer, an optimal solution must be chosen. These requirements have a direct relation to both the reliability and quality of the DC. But there is no approach to selecting ITIS for DC for a particular business case or to confirm that the existing ITIS is appropriate.

As a result of the analysis and the information presented above, for the assessment of the quality of the DC and the selection of the optimal option, are selected some of indicatorsdiscussed above and additional indicators.

A. Availability

Availability is one of indicators which used to assess DC's probability that it is operating correctly and is available to perform its function at the certain amount of time. Availability is typically used as ameasure of dependability for systems where shortinterruptions can be tolerated [3].

An availability factorK_{av} defined by the following:

$$K_{av}=\frac{MTBF}{MTBF+MTTR}*100\ [\%] \quad (3)$$

Availability of the DC is often specified in terms of downtime and uptime per year. If we use this terms the formula is[8]:

$$K_{av}=\frac{Uptime}{Uptime+Downtime}*100[\%] \quad (4)$$

Depending on the value obtained, one of availability classes (0 to 6) is assigned to DC (Table I). These classes usually are indicated by number of 9th.

TABLE I
AVAILABILITY CLASSES

Availability class	K_{av}, %	Number of 9th	Downtime
0	<90		
1	90	one	36.5 days/year
2	99	two	3.65 days/year
3	99,9	three	8.76 hours/year
4	99,99	four	52 min/year
5	99,999	five	5 min/year
6	99,9999	six	31 sec/year

Some authors asRohani, etc. [6], [9], classifyavailability as inherent, achieved and operational. The last of them is most important for us because is a measure of real average availability over a period of time and includes all experienced sources of downtime, such as administrative downtime, logistic downtime, etc [9].

B. RTO

As noted above, the RTO defines the time required for the full recovery of the DC operation after an accident (disaster, incident) reported from the moment of occurrence of the event. If we look at the time axis, the distance between the point where the incident occurs and the RTO point is a time interval during which applications are inaccessible, respectively business processes that are dependent on these applications are totally or partially discontinued. Now has a business processes with requirements this time to be a milliseconds.. Achieving such a break time is not always financially feasible and possible for many businesses in the world, but that does not prevent it from being desired.

On the other hand, any interruption of a business process results in damages that can be directly or indirectly financially assessed. Hence, assuming a calculation of the value of one minute, hour, or other time interval (most often a minute) that the business has determined, any interruption of work can be valued. This value is used as the basis on which the organization will determine the RTO and the IT solution that needs to be built in order to achieve the required RTO.

It is very important to specify that the achievement of a particular RTO is not just a matter of a technical solution. Recovery time depends on many components. One of the most important and obligatory components is the presence of Disaster Recovery Plan (DRP). The lack of such a plan makes the recovery of work unpredictable over time because the activities that will have to be carried out have yet to be clarified, coordinated with the management and carried out. The activities normally required to be performed when the work is restored are:problem analysis, damage assessment and order of needed equipment; installation of hardware and operating systems; installation of application software; restore data from latest backup; testing the system and troubleshooting and restoring work and announcing the end of the incident.Depending of ITIS of DC, this time can be very different and can be lower if implemented IT solution includes possibility to transfer operations to a spare center (cold, hot or hot.The timing of these activities depends on the competence of the staff and the periodic conduct of the exercises for the execution of the DRP.

The conclusion that can be drawn is that as much as the RTO is less, the price of the IT solution is higher.

C. RPO

As described in previous part,with RPO we denote a point in time before the incident to which the data will be recovered. This means that there will be lost data. The business must define how much data to be lost. An organization typically has information of a different type. Each type of information, during the work process, changes with different dynamics. It is therefore possible to reserve the different types of data in a different way in the same organization. For example, let's compare the change to the following two types of information: files and database in the chain of commercial stores. Even without accurate information, it can be very accurate to assume that loss of documents or change in documents for one hour will be much less significant than the loss of business data for the same time. Depending on the application software used in the latter case, work may not be easily recoverable in the absence of data, which means that such an IT solution must be implemented to prevent data loss.The RPO backup / replication strategy has a direct impact both the RTO size and the cost of the IT solution, i.e. the RPO is less, the higher the value of the IT solution.

Strategies that are used in practice, ranked in descending order of RPO, are: a backup of tape; a backup of a disk array; asynchronous replication of data; synchronous replication of data.If backups are stored in a location different than the backup centre, the RTO will increase with time for logistics.

D. CAPEX

The cost of building a DC is a complex indicator that consists of many components:

- cost of design IT solution (depended from RTO, RPO and K_{av});
- cost of hardware and software components of selected IT solution;
- cost of installation and implementation;

- cost of premises and engineering installations (not include in this paper).

CAPEX will be quantified using K$_{capex}$according Table II.

TABLE II
CAPEX CLASSES

K$_{capex}$	CAPEX (BGN)
1	>1 000 000
2	500 000 – 1 000 000
3	100 000 - 500 000
4	50 000 – 100 000
5	10 000 – 50 000
6	< 10 000

E. OPEX

Operational costs includes: salaries of IT personal or outsourcing IT services;license fees for system and application software;depreciation charges for of equipment and premises, maintenance of premises and all necessary engineering facilities (rents, electricity, cleaning, etc.). In this paper, these costs will not be taking into consideration.

OPEX will be quantified usingK$_{opex}$according Table III.

TABLE III
OPEX CLASSES

K$_{opex}$	OPEX(BGN)
1	>100 000
2	50 000 – 100 000
3	20 000 - 50 000
4	5 000 – 20 000
5	1 000 – 5 000
6	< 1 000

F. Implementation time

The implementation time is the sum of times for the following activities: logistics - $\boldsymbol{T_{log}}$;operating systems installation – $\boldsymbol{T_{os}}$;application software installation– $\boldsymbol{T_{app}}$; installation on premises – $\boldsymbol{T_{inst}}$;employee education - $\boldsymbol{T_{edu}}$.

$$T_{impl}= T_{log}+T_{os}+T_{app}+T_{inst}+T_{edu} \qquad (5)$$

IV. COMPLEX GENERALIZED INDICATOR

If describe customer requirements as follow:

$$CR=\{RTO, RPO, K_{AV}, K_{CAPEX,}K_{OPEX,}, T_{IMPL},\{IMPACT\}\} \quad (6)$$

where*Impact*is a vector that sets the weighting factors (importance) for each of the indicator. They are determined by the customer and may have one of the following values: 1 - low importance, 2 - medium importance and 3 - high importance.

The known complex arithmetic indicator K$_{CR}$can be used for quantification of the CR. K$_{CR}$ is calculating by formula:

$$K_{CR} = \sum_{i=1}^{6} b_i \, d_i \ , \qquad (7)$$

where$d_i$are standardized estimates of single indicatorsx_i(RTO, RPO, K$_{AV}$, K$_{CAPEX,}$K$_{OPEX,}$, T$_{IMPL}$), i.e. $0<[d_i = f(X_i)]\leq 1$, and b_i- the respective weighting factors.

$$\sum_{i=1}^{6} b_i = 1 \qquad (8)$$

The complex indicator K_{CR}provides an opportunity both for quantitative assessment of the quality of the DC and for the optimal choice between several variants of DC.

V. CONCLUSION

The quality of a Data Center (DC) can be assessed by different criteria. In this paperwas analyzed the indicators that can be used to assess the quality of DC from the point of view of the accessibility of the services provided. Based on the analysis was substantiated set of significant indicators for the DC quality. Quantification of the quality is offered using a complex generalized indicatorK_{CR}, which can be used to benchmark between different variants of DC and selects an optimal variant for concrete business case.

It was justified that there is a need for the use of a complex approach for assessing the sustainability of a HRDC.

ACKNOWLEDGEMENT

I would like to thank the anonymous reviewers for their valuable feedback.

REFERENCES

[1] Хр. Христов, В. Трифонов, *Надеждност и сигурностнакомуникациите*. София: Новизнания, pp. 15-25, 2005.

[2] A. Avizienis, J. Laprie, and B. Randell, “Fundamental Concepts of Dependability”, LAAS Rep. no. 01-145, 2014.

[3] E. Dubrova, “Fault-tolerant design”, *Fault-Tolerant Des.*, pp. 1-185, 2013. *Закон за защита прибедствия,* Държавен вестник, бр. 39, 2011.

[4] Cisco, “Design in Gand Managing High Availability IP Networks”, 2004.

[5] А. Павлов, “Структурный анализ катастрофоустойчивой информационной системы,” *СПИИРАН*, vol. 8, pp. 128–153, 2009.

[6] H. Rohani, and A.K. Roosta, “Calculating Total System Availability,” *Univ. van Amsterdam*, 2014.

[7] J. Gray, and D.P. Siewiorek, “High-Availability Computer Systems”, *Computer (Long. Beach. Calif)*, vol. 24, no. 9, pp. 39-48, 1991.

[8] ANSI/BICSI, *ANSI/Bicsi 002*. 2014.

[9] ReliaSoft Corporation, “Availability and the Different Ways to Calculate It.” [Online]. Available: http://www.weibull.com/hotwire/issue79/relbasics79.htm.