Journal of Polish Safety and Reliability Association
Summer Safety and Reliability Seminars, *Volume 8, Number 1, 2017*

Walkowiak Tomasz
ORCID ID: 0000-0002-7749-4251
Wroclaw University of Science and Technology, Faculty of Electronics, Poland

Marcin Pol
Wroclaw University of Science and Technology, Faculty of Computer Science and Management, Poland

The impact of administrator working hours on the reliability of the Centre of Language Technology

Keywords

Reliability, repair time, working hours, language technology infrastructure, natural language processing, CLARIN-PL

Abstract

The paper presents reliability analysis of CLARIN-PL Centre of Language Technology (CLT). The CLT is a Polish part of the language technology infrastructure developed by CLARIN project. The main goal of which is to support researchers in humanities and social sciences. The infrastructure is a complex computer system that enables combining language tools with language resources into processing chains. Authors present the system structure, analyse types of faults and define the CLT reliability model. The model takes into account the fact the time gap, between the failure and the repair of the system is not exponential since repair actions are taken only when administrators are at work (assuming that administrators are not working 24/7). The model is used to estimate reliability metrics (mean time and 90th percentile of relative down time and relative partial operational time) by a use of Monte-Carlo simulation. Moreover, analysis of possible improvement in the CLT organisation and its influence on the estimated metrics is given.

1. Introduction

CLARIN[1] (Common Language Resources and Technology Infrastructure) is a pan-European research infrastructure intended for the humanities and social sciences. Each member CLARIN is obliged to contribute parts of the language technology infrastructure [6]. In case of Polish part (CLARIN-PL[2]) it is operated by the CLARIN-PL Center of Language Technology (CLT) at Wroclaw University of Science and Technology. CLT is a computer system with Internet access. It provides a set of Web based applications with a user-friendly interface, which does not require advanced knowledge of information technology. It facilitates work with very large collections of texts. The CLARIN-PL CLT provides software, which enables the use of the previously developed digital archives and corpora, and new resources are created, stored and shared. It is also possible to work on raw text published on the Internet in such forms as press releases, articles, blogs and other documents. CLARIN-PL CLT users are researchers and students in the fields of humanities, social sciences and computer science. The main web page has in average 80 individual visitors per day. PL-WordNet [10], [11] web page has over 200 individual visitors per day. Some institutions like The Institute of Literary Research of the Polish Academy of Sciences are using CLARIN-PL resources [9] and tools [7], [8] every day.

For this reason, CLT reliability [1], [2], [3] is crucial. It is severely limited by the occurrence of faults, which degrade or compromise the CLT operation and the time required to perform the repair process.

One of the technique used to extend the system operation time when an incident occurs is to “isolate the affected system or network” and to “relocate the

[1] https://www.clarin.eu
[2] http://clarin-pl.eu

target" [12]. Relocating of services (reconfiguration [4], [13]) is always connected with some period when they are not available to the end-users.
Due to the random nature of incidents occurring in computer systems (and CLT is a computer system) stochastic processes are used for mathematical modelling of the system reliability. The most often used stochastic processes in technical system reliability analysis are both Markov and semi-Markov processes [1],[3]. Markov processes are limited to exponential distributions of all state transitions. This is a reasonable simplification in case of fault occurrences, but it is questionable in case of the repair processes, especially in case of the service relocations, that requires human administrators. In case of CLT, system administrators are not working in the three-shift system (24/7) [14]. Therefore, there are periods when no administrator is at work. We propose to use a realistic model of a repair time [15],[16] that takes into consideration working hours of administrators. To overcome limitations of Markov model, the presented analysis is performed using Monte-Carlo approach [5]. Achieved numerical results are compared with results achieved assuming Markov approach.
The paper is structured as follows. In the next section, we present CLARIN-PL Centre of Language Technology systems. It is followed by analysis of faults in CLT. Next, the S-T model of CLT is defined. It is followed by a short presentation of reliability analysis method, which uses repair time model with administrator working hours. Finally, the method is applied to S-T model of CLT and a set of reliability metrics is defined. Moreover, the three scenarios of changing the CLT organisation (changing working hours and/or upgrading UPS system) are analysed and improvement of the metrics are presented.

2. CLARIN-PL Centre of Language Technology

CLARIN-PL is a Polish research consortium, a section of pan-European research infrastructure – CLARIN. It consists of six scientific units, where electronic language resources and tools for working with large collections of texts in Polish are being created and developed.
CLARIN infrastructure is a network of centers whose tasks are provide users with tools and resources related to natural language processing – type B[3]. Polish CLARIN node – CLARIN-PL CLT – provides users with tools and resources related to natural language processing (type B) is being built at Wrocław University of Science and Technology. Thanks to the strict observance of accepted standards, users registered in the CLT are granted free access to tools and language resources available both in Poland and in CLARIN centers in other member states.
The aim of the CLT is filling gaps in the system of Polish basic tools and resources. It actively cooperates with scientists in the fields of humanities and social sciences in order to create and develop innovative, e-humanities-oriented applications for the Polish language. Ultimately, cooperation may also include digital libraries, archives, museums, etc.
CLARIN tools support a variety of language-processing tasks, including automatic text summarization, search for entity names, and morphological and syntactic analysis of documents. Processing of this kind will help, for example, those who study the political, social or advertising discourse.
CLT tasks includes:

- maintain the repository where tools are collected and resources are labelled with permanent identifiers;
- taking care of technical coherence of the system and its compliance with the standards, intellectual property rights, licenses, and ethical rules;
- maintain a security policy, for example, through certification of servers and responsible management of personal data.

To maintain all tasks a private cloud was installed. Hardware consists of nine servers in a mixed rack / blade architecture. Each server has 192 GB to 224 GB of RAM. Which gives a total 1967 GB of RAM (~ 2 TB). Each server has two Intel (R) Xeon (R) CPUs E5-2665@2.40GHz, which let you run up to 16 threads per processor. In total, it gives power 324 processes in parallel.
Data storage subsystem is built on IBM Storwize V7000 with redundant dual-active intelligent FC 8 Gb controllers and dual-active iSCSI controllers. Storage is using RAID10 volumes (Redundant Array of Independent Disks; in this mode each chunk of data is repeated). FC 8 Gb controllers are connected by redundant Fiber Interconnects with NEXUS switch and vNIC Ethernet cards to convergence VIC 1227 blade servers cards. iSCSI ports are connected to VIC 1227 10 Gb cards (installed in C220 server) with iSCSI ports on data storage V7000 controllers.
Data transfer is carried out using fiber optic cabling, which is the key element of the data rate. Data backup is implemented on DS3500 Storwize V7000, ProtecTIER 6710 IBM System with deduplication mode. System is configure to create complete data snapshot every Sunday. All system is protected by UPS system.

[3] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-78

To maintain this expectations additional server module has been purchased to provide High Availability (HA) and data replication. The additional tenth server has its own separate disk storage. The storage is also using RAID10. The above solution enables performance in data flow that is significant when processing files of the order of 1TB.

CLT use complete automation tool for managing Xen server[4] pools, which utilize the XAPI management interface and toolstack. The software suite provides complete HA features within a given pool. The overall design is intended to be lightweight with no compromise of system stability. HA is provided with built in logic for detecting and recovering failed services.

CLARIN-PL server have two virtual machine servers, with failover, provide safe environment to run services. Service is defined as the application and underlying operating system.

CLT provide scripts for Xen server with features of:

- auto-start of any failed VMs,
- auto-start of any VMs on after reboot,
- detection of failed hosts and automated recovery of any affected VMs,
- detect and clean up orphaned resources after a failed host is removed,
- removal of any failed hosts from pool with takeover of services,
- reconfiguration virtual hosts,
- movement virtual hosts to other servers.

3. Fault analysis

When considering the reliability of CLT system, it is necessary to analyse a very diverse set of faults. It encompasses energy power down, hardware faults, errors in the software or security vulnerabilities.

The most suitable for the proposed analysis is the analysis of faults that is based on the effects they have on the CLT system and its users. The proposed taxonomy of faults, as described in *Table 1*. The “x” symbol in column means that fault described in row need admin or external support.

[4] https://www.citrix.com/products/xenserver/

Table 1. Taxonomy of faults

Faults	**Need CTL admin support**	**Need external support**	**Operation state**
Hardware faults:			
Complete storage failure	x	x	no
Single disk error	x		part
Storage Controller	x	x	part
RAM memory	x	x	part
Processor failure	x	x	part
Mother board	x	x	part
Ethernet Switch	x	x	part
UPS failure	x	x	no
Wiring	x	x	part
Air condition		x	no
Errors in the software:			
Virtual machine error	x		part
Virtual machine host down	x		no
Software/services error	x		part
Power down:			
Server Power supply	x	x	part
Lack of energy (below UPS power up time)			full
Long lasting lack of energy	x		no
Security vulnerabilities:			
Lack of Internet		x	no
DDoS attacks	x		part

4. Reliability model of CLT

4.1 Levels of operation and types of repairs

From the point of end-user, we propose to distinguished three different levels of operation of the CLT system:

- fully operational,
- partially operational – the system is available to the end user, but the performance is degraded,
- inoperational.

Table 2. Mapping between type of faults (*Table 1*) and ways of repair

Faults	Ways of repair process
Hardware faults:	
Complete storage failure	reconfiguration
Single disk error	administrator
Storage Controller	administrator
RAM memory	administrator
Processor failure	administrator
Mother board	administrator
Ethernet Switch	administrator
UPS failure	reconfiguration
Wiring	administrator
Air condition	reconfiguration
Errors in the software:	
Virtual machine error	administrator
Virtual machine host down	administrator
Software/services error	administrator
Power down:	
Server Power supply	reconfiguration
Lack of energy (below UPS power up time)	automatic
Long lasting lack of energy	conditional
Security vulnerabilities:	
Lack of Internet	automatic
DDoS attacks	administrator

Following types of faults presented in the previous system and the need of CLT administration intervention we could distinguished four different ways of repair process:

- automatic – repair is done independently from CLT and after the repair the system if fully operational (for example network failure),
- administrator – repair is done by CLT administrators and after the repair the system if fully operational (for example restarting the server/service),
- reconfiguration – after the failure system is inoperational, then it is reconfigured by the administrators and is partially operational, it if followed by the repair and the system if fully operational,
- conditional – repair is done independently from CLT but depending of the length of the failure the system if fully operational or (for longer failures) partially operational and requires CLT administrators intervention to become fully operational.

Mapping between fault types analysed in the previous chapter level operation and ways of repairs is presented in .

4. 2 State transition model

Considering types of repairs presented in the previous chapter following reliability states of CLT could be distinguished:

- normal operation (S_0),
- inoperation in case of conditional repair (S_1) – it models the situation of power down in building where CLT is located and the situation that modern system has UPS but they have a limited capacity so there is a probability that a cut off of electricity will cause the system down; the intensity of power downs is marked by λ_P, the repair intensity is marked by μ_P, power repair is not performed by web system administrator with limited working hours so it will be modelled by exponential distribution,
- after power down (S_2) – the system is inoperational till some administrators actions will be taken, intensities of repair is equal to: δ_S,
- failure of network resulting in system inoperation (S_3) – with failure intensities λ_N and intensities of repair: δ_N,
- software or hardware failure (S_4) – the short/small software, hardware failure results in system inoperation, it occurs with intensity: λ_H; after the failure administrator actions are performed (intensity of repairs δ_H),
- hardware failure (S_5) – the hardware failure witch needs external support like hardware producer support; the system is inoperational till administrators actions will be taken and run some of services on backup hardware (S_6); it results in system inoperation, it occurs with intensity: λ_R ;after the failure administrator actions are performed (intensity of repairs δ_{RR}),
- after hardware failure with needs external support, running services on backup hardware (S_6) – the system is half-operational/slow- operational till administrators and external support actions are performed (results in system not fully operation,

it occurs with intensity: δ_{RR}; after the failure administrator actions are performed (intensity of repairs δ_R).

The S-T model is presented in *Figure 2*.

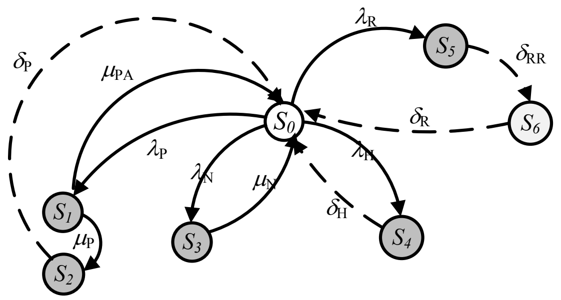


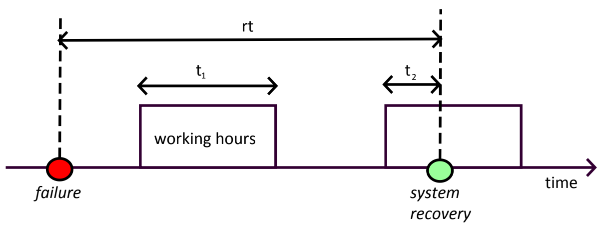
Figure 1. The S-T model of CLT system

All repair processes with intensities marked by δ and dashed lines requires CLT administrators. In S_0 system is operational, in S_6 partially operational, in other states (dark filled in *Figure 1*) is inoperational.

5. Repair time model

The transitions in S-T model (*Figure 1*) marked by dashed line represents the repairs, which requires administrators work. Since administrators in CLT and not working in 24h manner the repair time (*rt*, time from failure occurrence to restoring the system to operation) includes not only the work time of the administrator (real repair time - *rrt*) but also a time when the administrator is not working (weekends, nights).

Since, working hours depends on a time therefore the repair time is a function of time when failure occurs and real repair time. It could be calculated using Monte-Carlo simulation [5] following algorithm presented in [16].



$rrt=t_1+t_2<=rt$

Figure 2. Relation between real repair time (*rrt*) and repair time (*rt*)

During numerical experiments, the results of which are presented within this paper, we have assumed that working hours of administrator are 8 am to 16 pm, Monday to Friday unless it is stated that other results were used (scenario 2 and 3 in chapter 6.3).

6. Experiments and results

6.1. Reliability parameters

We have assumed following values of reliability parameters:

- λ_P = 2/year
  (power failure intensity),
- $1/\mu_{PA}$ = 0.5 h
  (power down repair mean time that results in automatic return to operational state, it is equal to UPS operating time),
- $1/\mu_P$ = 4 h
  (power down repair time that will require administrators intervention),
- $1/\delta_P = x + 2$ h
  (administrators intervention *rrt* after power down)
- λ_N = 4/year
  (network failure intensity),
- $1/\mu_N$ 2 h
  (network repair time),
- λ_H = 2/year,
  (hardware or software failure intensity)
- $1/\delta_H = x + 2$ h
- λ_R = 2 per 3 years
  (failures that requires reconfiguration of the system)
- $1/\delta_{RR} = x + 2$ h
  (reconfiguration time),
- $1/\delta_R = x + 16$ h

All repairs times that requires administrators works are marked as *x* + some time. Where *x* indicates the difference between repair time and real repair time.

6.2. Reliability metrics

Reliability of the system could be described by different metrics. From the user point of view the most import (and often used in service level agreements for internet services) is a down time presented as a percentage value. It is calculated by dividing the sum of all downtimes timespans by a analyzing time span. It is important to state that such define time is a random value. Moreover, in the analyzed system (as mentioned in 4.1) we have three levels of operation that is why we propose to analyze two downtimes (random variables):

- relative down time (RDT) - sum of times when the CLT system on the inoperational level divided by analyzed period,
- relative partial operational time (RPOT) - sum of times when the CLT system is on the partial operational level divided by analyzed period,

and their statistics:

- mean value (marked as *mrdt* and *mrpot* respectively),
- 90th percentile (marked as *90rdt* and *90rpot* respectively) i.e.:

$$P(\text{RDT} < 90rdt) = 0.9 \qquad (1)$$

The 90th percentile could be understood as a guaranteed (with 0.9 probability) relative down time. Such statistics are justified from economic point view.

6.3. Results and comparative analysis

The reliability statistics proposed above were calculated for realistic repair time model with truncated Gaussian model of real repair time and working hours set to 8 am to 4 pm using Monte-Carlo based simulator developed by authors [8]. Results are presented in *Table 3.*

Table 3. Reliability metrics

Statistics	*Mrdt*	*90rdt*	*mrpot*	*90rpot*
Values	1.321%	2.155%	0.449%	1.086%

Next, we have analyzed the CLT system reliability model for real repair time driven by exponential distribution. Relative differences (in percentage) to the truncated Gaussian real repair times (*Table 3*) are presented in the second row of *Table 4*.
It could noticed that the results are very similar except the 90*rpot* metric. It is caused by the fact that time of being the system in S_6 (partially operational) strongly depends on the repair time that involves administrator intervention (transition in S-T model marked as δ_{RR}, δ_R in *Figure 1*). The exponential distribution has much longer right tail than Gaussian one, and it causes larger values of 90*rpot.*
In the next step, we used Markov assumptions to solve the S-T CLT system model (*Figure 1*). The mean values of repair times that requires administrators were enlarged according to the method presented in [16] to include the difference between repair time and real repair time (see *Figure 2*).
Relative differences (in percentage) to the base model (results in *Table 3*) are presented in the third row of Tab. 4. As it could noticed the exponential distribution for all sate change times gives much larger (from 12% to 55%) estimators of analyzed reliability metrics.

Table 4. Relative differences to results in *Table 3* for models with exponential real repair time and Markov model

Statistics	*mrdt*	*90rdt*	*mrpot*	*90rpot*
Simulation with exponential *rrt*	1.11%	2.83%	-1.9%	14.0%
Markov model	11.79%	15.02%	39.83%	54.6%

6.4. Changes in CLT system organisation

The presented method allows calculating values of reliability metrics for different values of reliability model parameters. It could be used to justify changes in the CLT system organization. Let us that we are analyzing three scenarios of the CLT system reliability improvement:

- Scenario 1: adding additional UPS, i.e. $1/\mu_{PA}$ to 4 h,
- Scenario 2: hiring additional administration staff and two shift work, i.e. changing working hour to 6 am - 8 pm,
- Scenario 3: both changes from scenario 1 and 2.

Table 5. Relative improvement in reliability metrics for different scenarios

Metric	*mrdt*	*90rdt*	*mrpot*	*90rpot*
Scenario 1	18%	14%	0%	0%
Scenario 2	23%	21%	43%	46%
Scenario 3	36%	32%	43%	46%

The results are presented in *Table 5*. For example adding additional UPS (enlarging UPS operating time to 4 hours) will give 16% shorter mean down time. Calculated values and assessed costs (for each scenario) will to justify decisions on changing the CLT system organization.

7. Conclusion

The paper presents a reliability analysis of CLARIN-PL Centre of Language Technology (CLT). We model the system reliability by S-T model with realistic model of repair time i.e. all repairs that requires human intervention are taken only during working hours of administrators. The model is solved by Monte-Carlo simulation and it allows calculating such metrics as mean time and 90th percentile of relative down time and relative partial operational time.

The presented method and the developed tool allow exploring the impact of changes in the system maintenance (such as working hours of administrator or UPS capacity) on the CLT system. Operators of any complex computer system that has to operate 24/7 may use the proposed method.

References

[1] Aven, T., Uwe, J. (1999). Stochastic Models in Reliability. New York: Springer.

[2] Avizienis, A., Laprie, J. , Randell, B. (2000). Fundamental concepts of dependability. In Proc. 3rd IEEE Information Survivability Workshop, Boston, Massachusetts, pp. 7–12

[3] Barlow, R. & Proschan, F. (1996). *Mathematical Theory of Reliability*. Society for Industrial and Applied Mathematics, Philadelphia.

[4] Caban, D. & Walkowiak T. (2012). Preserving continuity of services exposed to security incidents. *Proc. The Sixth International Conference on Emerging Security Information, Systems and Technologies*, SECURWARE 2012, IARIA, 72–78.

[5] Fishman, G. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag.

[6] Hinrichs, E. & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In ed. N. Calzolari et al., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ELRA*

[7] Maryl, M. & Piasecki, M. (2015). Text Clustering Methods in Literary Analysis of Weblog Genres. *Expert Workshop on Topic Models and Corpus Analysis, DARIAH Text & Data Analytics Working Group, Dublin* 14-15.12.2015.

[8] Maryl, M., Piasecki, M. & Młynarczyk, K. (2016.) From interpretation to corpus stylistics and back. Towards an integrated model of the emergent genres classification. *15th conference of International Association for Empirical Study of Literature and Media (IGEL), Northern Illinois University, Chicago*, 6-9.07.2016.

[9] Maryl, M., Piasecki, M. & Młynarczyk, K. (2016), Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres. *Digital Humanities 2016, Uniwersytet Jagielloński i Uniwersytet Pedagogiczny, Kraków*, 11-16.07.2016.

[10] Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S. & Kędzia, P. (2016). plWordNet 3.0 -- a Comprehensive Lexical-Semantic Resource. *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, red. Calzolari, N.; Matsumoto, Y., Prasad, R., December 11-16, 2016, Osaka, Japan*, ACL, 2016, s. 2259–2268

[11] Piasecki, M., Burdka, Ł., Maziarz, M. & Kaliński, M. (2016). Diagnostic Tools in plWordNet Development Process. *Human Language Technology. Challenges for Computer Science and Linguistics, red. Vetulani, Z.; Uszkoreit, H. & Kubis, M., Springer,* LNCS 9561, s. 255–273

[12] Scarfone, K., Grance, T. & Masone, K. (2008). Computer Security Incident Handling Guide, rev. 1, *NIST Special Publication 800-61*

[13] Walkowiak, T. & Caban, D. (2015). Improvement of dependability of complex web based systems by service reconfiguration. *Dependability Problems of Complex Information Systems*. Advances in Intelligent Systems and Computing 307, 149–165.

[14] Walkowiak, T. (2014). Simulation based availability assessment of services provided by web applications with realistic repair time. *Eksploatacja i Niezawodnosc – Maintenance and Reliability*. 16 (2), 341–346.

[15] Walkowiak, T. (2014). Weekly availability changes of web applications implemented in virtualized environment. Nowakowski, T. et al. *Safety and Reliability: Methodology and Applications,* CRC Press.

[16] Walkowiak, T. (2016). A repair time model of a web based system including administrator working hours. *Journal of Polish Safety and Reliability Association*, 7(1), 233-240.