# Simulation of Information Lifecycle Management

Lars Arne Turczyk, TU Darmstadt, KOM Multimedia, Communications Lab, Merckstr. 25, 64283 Darmstadt, Germany; E-mail: lars.turczyk@siemens.com
Oliver Heckmann, TU Darmstadt, KOM Multimedia, Communications Lab, Merckstr. 25, 64283 Darmstadt, Germany
Ralf Steinmetz, TU Darmstadt, KOM Multimedia, Communications Lab, Merckstr. 25, 64283 Darmstadt, Germany

## ABSTRACT

*In this paper we analyze the effects of the number of storage hierarchies in an ILM system. We describe the model for our simulator used to run the simulations. Afterwards the results are compared and recommendations are made.*

**Keywords:** Information Lifecycle Management, ILM scenarios, storage hierarchies

## 1. INTRODUCTION

Information Lifecycle Management (ILM) is a strategic concept for storage of information and documents in which the value of the stored objects depends on the underlying business model and processes. Documents are assigned to a storage medium automatically so that the existing storage capacities can be used optimally and more cost efficiently.

For realizing cost potentials it is necessary to obtain a wider knowledge about ILM procedures and scenarios but experience reports do not exist in sufficient form and experimenting and searching in real systems is too expensive. Therefore the aim of this paper is to generate results and experiences by simulation of ILM scenarios.

First we work out the aims of a simulation and then create the corresponding simulation model. The model allows a strategy-orientated analysis. Based on this model a simulator was implemented as an examination tool for the behavior of ILM scenarios. The simulation model uses results of our study conducted in 2006 [1] which analyzed the access behavior on documents of a company database. The study provided a statistical description of the access patterns which is used in the model for the automatic migration of files.

The scientific use of the work consists of providing a model for simulation which generates generally utilizable results concerning ILM behavior and cost optimization.

The results focus on the optimal number of storage hierarchies in an ILM system

The paper starts by listing the objectives of ILM simulations and describing the simulation model. Then the scenarios to be simulated are defined. Simulation results are presented and interpreted. The paper ends with an outlook on further ILM simulations.

## 2. RELATED WORK

Strange examined the long-term access behavior on files in an UNIX system [2]. His aim was to identify regularities and patterns which can be applied to automated migration strategies for Hierarchical Storage Management (HSM). To verify hypotheses on migration algorithms a simulator was also designed and implemented. His examination is different to our approach on implementation. The simulator developed by Strange served as a tool merely for checking migration algorithms which were verified using observed access behavior. A stochastic simulation of the access behavior was renounced. Instead, the user behavior was generated deterministically from the access protocols. Since only the effect of the migration rules was analyzed, he could restrict the number of feigned storage hierarchies to two. In addition, only very simple migration algorithms were used.

Further work deals mainly with algorithms which can be used for ILM or other storage strategies. Some examinations focus on the analysis of the access behavior and the development of migration strategies. The file migration protocol listing of a supercomputer was analyzed in a study by Miller and Katz. Migration methods were developed for a corresponding system [3].

Schmitz has also analyzed the access behavior on files in a supercomputer to be able to derive an optimal migration strategy [4].

Miller and Gibson examined the access behavior in further studies in UNIX environments and designed a “file aging algorithm” as a migration rule [5].

Today ILM is a strict focus of research. The main results are found on the field of “How” ILM works, i.e. most research was done on the field of procedures and policies. Vendors gave their point of view about ILM understanding [6] Turczyk et. al. gave a formal definition usable for ILM abstraction [7]. Chen focused on the valuation of files [8]. Proposals for policy description of ILM were presented by Beigi et. al. [9] and Tanaka et. al. [10]. Beigi et. al. considered the file system environment. Tanaka et. al. rules are more general and offered a time schedule for migration. Both papers intend to use metadata for ILM realisations.

Our paper is influenced by the analysis of the long-term access behavior of Turczyk et. al. [1]. In contrast to other work, they analyzed Microsoft office files of a company database. In addition, they derived complex statistical rules for the migration of office files and documents.

## 3. SIMULATION MODEL

Our objectives below list how to implement a simulator as an examination tool for ILM scenarios. The primary objectives are the analysis of fundamental questions of ILM:

- Integral analysis of ILM scenarios (end-to-end)
- Identification of the necessary number of storage hierarchies

The simulator takes into consideration the integral lifecycle, i.e. from the initial situation designed for a company to the point where a stable state is reached. The simulator should offer transferable results concerning the questions mentioned above.

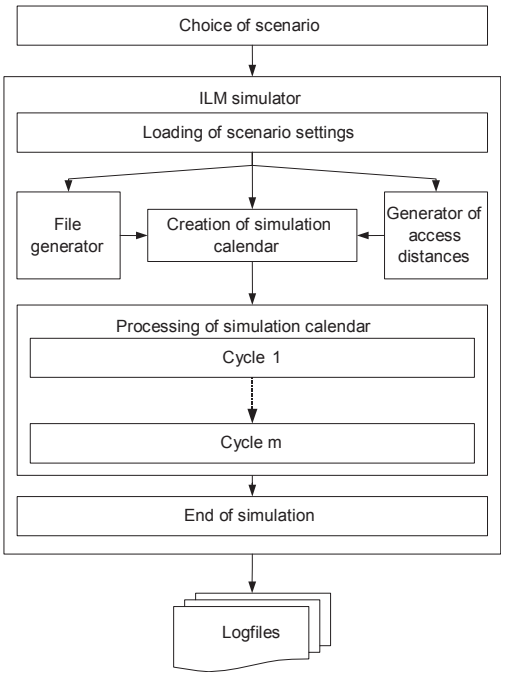
Some aspects of ILM have not been the subject of focus with this simulator and must therefore be considered separately. These secondary objectives are, in particular,:

- Identification of the necessary number of migration rules
- Optimization of the wording of migration rules
- Analysis of the dynamic behavior of ILM scenarios
- Realization of potential for cost reduction

Figure 1 shows the simulation model with its structural layout:

The main component of this plan is the ILM simulator. A scenario is loaded and simulations are executed. As a result the simulator generates logfiles. Any evaluation and interpretation of the results are done externally. How the simulator works is shown in the next section.

*Figure 1. Simulation model*




## 4. SIMULATIONS

For the examination of the effect of the number of storage hierarchies four simulation runs are carried out. The assumed data growth is about 20% per annum. The simulator starts the simulation with a data stock of 500 files. The simulation duration is 2,000 days. Ten simulation runs are averaged to one simulation to reduce fluctuations of measurements.

The migration rules used in the simulations are based on our study [1]. As distribution function either the Weibull-distribution (W(α;β)) or Gamma-distribution (G(α;β)) is used (see table 1).

The number of storage hierarchies is the initial variable of the simulations. At every simulation the first level has a threshold probability of p1=10%. The distances of the threshold probabilities dpi of the other levels i are equidistant, i.e. the remaining 10% are split up equally (see figure 2).

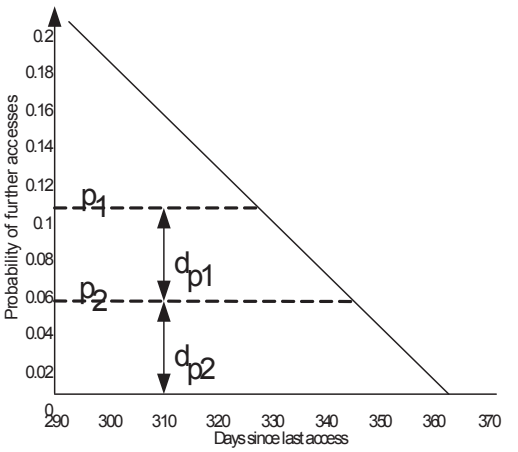
The threshold probability is described as the probability of further accesses on the file, where the file is assigned to a new hierarchy. In the example shown the probabilities of the first and second level are p1=10% or rather p2=5%. When the probability of further accesses on a file stored on hierarchy 1 falls below the threshold probability of 5%, it is migrated to hierarchy 2.

When adding a new storage hierarchy the threshold probabilities are adapted correspondingly so that they are of equal distance to each other again.

*Table 1. Applied distribution functions*

| Number of accesses / File type | 1-6 | 7-14 | 15-∞ |
|---|---|---|---|
| doc | W(0,35;3,5) | G(0,32;183) | W(0,35;3,5) |
| xls | W(0,25;1,1) | W(0,25;1,1) | W(0,25;1,1) |
| ppt | W(0,38;14,3) | W(0,38;14,3) | W(0,38;14,3) |
| pdf | W(0,35;3,5) | G(0,32;183) | W(0,35;3,5) |
| other | W(0,46;27,7) | G(0,29;181) | W(0,46;27,7) |

*Figure 2. Equidistant threshold probabilities in case of two hierarchies with p1=10% and p2=5%*




The influence on the number of hierarchies is observed by means of the relative capacity-need. In addition the jitter serves as a measure to look at the reliability of the system.

Now the individual simulation-runs and the accompanying results are explained.

At the first simulation there is only one threshold probability of p1=10% which lies between level 1 and 2. Figure 3 shows the result of the simulation.

In simulation 1 the relation between hierarchies 1 and 2 is approximately 1:1, i.e. almost half of the complete data stock is stored on the second, more economical hierarchy level. The average jitter is J(1000)=2.136.

*Figure 3. Mean relative capacity-needs for two hierarchies*

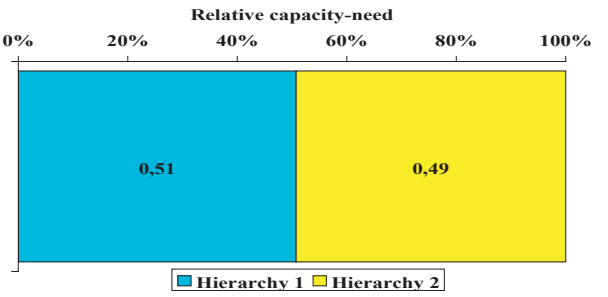
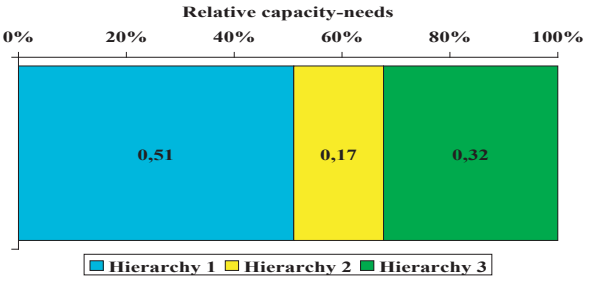



*Figure 4. Mean relative capacity-needs for three hierarchies*



In simulation 2 three hierarchies are available for the storage of the files. The related value probabilities are p1=10% and p2=5 %.

Figure 4 represents the result of simulation graphically by means of the relative capacity need.

Again approximately 50% of the data are on the first storage hierarchy. On the second hierarchy nearly a sixth of the complete stock is kept and on the third hierarchy nearly a third of the files is stored.

A mean jitter of J(1000)=2.093 was measured.

In simulation 3 the probability values are p1=10%, p2=6.66% and p3=3.33%. The mean capacity values arising from the simulation are represented in figure 5.

Hierarchy 1 keeps 51% and hierarchy 2 keeps 10%. Hierarchies 3 and 4 keep 14% and. 25% respectively.

The average jitter is J(1000)=2.16.

*Figure 5. Mean relative capacity-needs for four hierarchies*

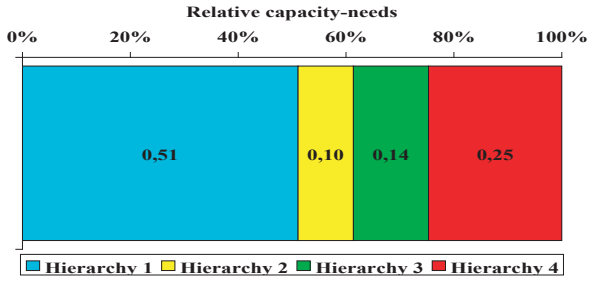



*Figure 6. Mean relative capacity-needs for five hierarchies*

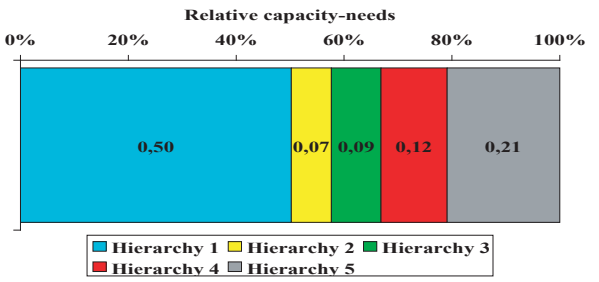
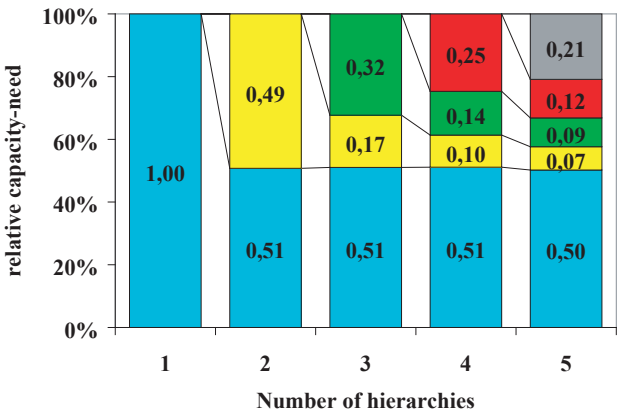



*Figure 7. Overview of the mean capacity-needs*




In simulation 4 a scenario with 5 hierarchies is simulated. The related probability values are p1=10%, p2=7.5%, p3=5% and p4=2.5%.

The results are shown in figure 6. Hierarchy 1 keeps 51 % and hierarchy 2 keeps 7.4%. Hierarchies 3 and 4 keep 9.2% and 12.3% respectively. Hierarchy 5 keeps 20.9%.

The measured jitter was J(1000)=2.17.

## 5. RESULTS AND INTERPRETATION

The observed jitter values of the four simulations vary less than 5%. It can be assumed that the reliability of an ILM system is independent of the number of hierarchies.

Figure 7 gives an overview of the capacity-need of the different scenarios.

By applying the sensitivity analysis we examine the effects of a variation in the number of hierarchies.

The storage demand of the first hierarchy remains constant irrespective of changes in the number of storage hierarchies.

The capacity need of the second level is reduced with an increasing number of storage hierarchies.

The reason for this behavior of the ILM system is the specification of the threshold probabilities.

With any added hierarchy the distances of the values change.

If there are only two hierarchies, the second hierarchy stores all files with an access probability of less than 10%.

If three hierarchies are used, the second only keeps files with an access probability of between 5% and 10%. Therefore the relative capacity need of the second level becomes smaller.

Generally speaking from all hierarchies lower than hierarchy 1, the lowest hierarchy always keeps the largest part of the files. Altogether, the greatest share of the information is stored on the top and the bottom hierarchies. This result coincides with the observations in real IT systems and is an essential driver for ILM. [11].

## 6. SUMMARY AND OUTLOOK

We presented simulation results for Information Lifecycle Management. The objective was focused on the optimal number of storage hierarchies. Although the number depends on the definite business process, the range of numbers of hierarchies could be isolated. In the next step further ILM scenarios will be simulated and compared. The focus will lie on the secondary objectives listed in section 3.

## REFERENCES

[1] Turczyk, Lars.; Groepl, Marcel; Heckmann, Oliver and Steinmetz, Ralf: Analyse von Datei-Zugriffen zur Potentialermittlung fuer Information Lifecycle Management, TU Darmstadt KOM Technical Report 01/2006

[2] Strange, Stephen: Analysis of Long-Term UNIX File Access Patterns for Application to Automatic File Migration Strategies. Technical Report UCB/CSD-92-700, EECS Department, University of California, Berkeley, 1992.

[3] Miller, Ethan L. and Katz, Randy H.: An Analysis of File Migration in a UNIX Supercomputing Environment. USENIX Winter, pages 421–434, 1993.

[4] Schmitz, Carolin: Entwicklung einer optimalen Migrationsstrategie für ein hierarchisches Datenmanagement System. Technischer Bericht, Forschungszentrum Jülich, 2004.

[5] Gibson, T. and Miller E.: An Improved Long-Term File-Usage Prediction Algorithm, 1999.

[6] Reiner, David; Press, Gil; Lenaghan, Mike; Barta, David; Urmston, Rich; "Information Lifecycle Management: The EMC Perspective" 20th International Conference on Data Engineering (ICDE'04) 1063-6382/04

[7] Turczyk, Lars Arne; Heckmann Oliver; Berbner, Rainer and Steinmetz, Ralf: A Formal Approach to Information Lifecycle Management. In Proceedings of IRMA ´05, Washington DC, May 2005

[8] Chen, Y.: Information valuation for Information Lifecycle Management. International Conference on Autonomic Computing (ICAC'05), 2005.

[9] Beigi, Mandis; Devarakonda, Murthy; Jain, Rohit; Kaplan, Marc; Pease, David; Rubas, Jim; Sharma, Upendra; Verma, Akshat, "Policy-Based Information

Lifecycle Management in a Large-Scale File System", IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'05)

[10] Tanaka, Tetsuo; Ueda, Ryoichi; Aizono, Toshiko; Ushimjima, Kazutomo; Naizoh, Ichiro; Komoda, Norihisa; Proposal and Evaluation of Policy Description for Information Lifecycle Management. International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05) 0-7695-2504-0/05

[11] Gostner, Roswitha; Turczyk, Lars; Heckmann, Oliver; Steinmetz, Ralf: Analyse von Datei-Zugriffen zur Potentialermittlung fuer Information Lifecycle Management, TU Darmstadt TR01/2005