IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB12755

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

An Analytical Model of Information Lifecycle Management

Lars Arne Turczyk, Oliver Heckmann, Rainer Berbner, & Ralf Steinmetz
TU Darmstadt, KOM Multimedia Communication Lab, Merckstr. 25, 64283 Darmstadt, Germany, P: +49697972778, F: +49697973437,
lars.turczyk@siemens.com

ABSTRACT

In this paper we derive an analytical model for Information Lifecycle Management (ILM) systems and use that model to investigate the cost saving potential of ILM systems and to support decision finding. We show analytically that if ILM is employed correctly it can lead to significant storage costs savings in enterprises.

INTRODUCTION

ILM is based on the idea that in an enterprise there are different information with different values. The different information will be stored on different storage devices.

ILM manages information according to its value. Valuable information is stored on systems with high Quality of Service (QoS). The value changes over time and therefore migration of information to cheaper storage systems with lower QoS is required. Automated migration makes ILM dynamic.

In this paper we identify the cost factors and their influences. The paper is structured as follows: Section 2 introduces the analytical model whose implications are investigated in section 3. At the end, in section 4, we show the applicability of the model and how storage decision finding is supported.

ANALYTICAL MODEL OF ILM

In this chapter an analytical model based on analyst studies is derived. First we check if the effects of data growth are neutralized by price declining (or performance improvement) of storage components. Of course the well-known Moore's law [1] and similar predictions can be applied for deriving the model. Here analysts' statements are used to take the current market into account to provide more specific predictions. The following table summarizes all used abbreviations in order to support the reading of the paper.

Table 1. Abbreviations

$g, g(t)$	growth rate of capacity demand
$g_i(t)$	growth rate of capacity demand in hierarchy i
$d, d(t)$	price decline per GB
$d_i(t)$	price decline per GB in hierarchy i
$c(t)$	cost
${}^n c(t)$	n-dimensional cost
$a(t)$	total amount of needed capacity
$a_i(t)$	amount of needed capacity in hierarchy i
$p(t)$	price of needed capacity
$p_i(t)$	price of needed capacity in hierarchy i
$ManC$	managing cost
α_i	hierarchy i's fraction of the total amount of needed capacity
β_i	price factor between hierarchy i and hierarchy 1 with $\beta_1 = 1$ always
${}^n r = \frac{{}^n c(t)}{{}^1 c(t)}$	ratio between n-dim.cost and 1-dim. cost

Data Growth

The University of California concluded that in 2002 alone around 5 exabytes (10exp18 Bytes) of new "stored information" were produced [2]. Even more amounts of data will be produced over the next few years, several analysts report. They all speak of steadily growing capacity demand. The compound annual growth rate (CAGR) varies between 60% and 100%. Metagroup identified a growth rate of 60% [3]. In 2001 IDC estimated a CAGR of 76% over the years 2000-2004 [4]. A report created in 2003 by Horison speaks of a CAGR for data demands over the next few years of 60%-70% [5]. Although the exact value is not known the tendency is obvious. To conclude we assume the capacity demand grows by a factor g ε [0.6; 1.0] per year.

Price Decline of Hardware

As the tendency for demand is growing the tendency for storage prices is declining. Again analysts give a range of prognoses. Between 1998 and 2001 McKinsey determined for the price pergigabyte (GB) a CAGR of -36% [6]. IDC took a look at the prices per GB between 2001 and 2003. In 2003 per-gigabyte external storage prices fell -33%, while in 2002 and 2001, they fell down -40% and -43% respectively. So the CAGR between 2001 and 2003 is -36% [4]. To conclude we assume the prices per GB decline realistically by a factor ε [-0.33;-0.36] per year.

Managing Cost

The cost situation changes when considering the total cost of ownership (TCO). Both Gartner and IDC reported that an enterprise spends an average of $3 managing storage for every $1 spent on hardware [5]. Additionally Gartner Group speaks of $3.5 being spent for managing each $1 spent for storage hardware [7]. This relation of approximately 3:1 between managing cost (ManC) and hardware cost has to be considered in the cost model, too.

n-dimensional Cost Model

ILM takes into account that in enterprises a lot of unused data is stored on high performance storage devices [8, 9, 10]. ILM assumes that the storage environment employs different hierarchies. Hence the cost model for ILM has to be extended to reflect the n-dimensional characteristic of an ILM solution.

When information are migrated between different hierarchies the cost per each hierarchy has to be considered.

Definition 1 (Multidimensional Cost) The TCO for an n-dimensional ILM solution is:

$$^{n}C(t) := \sum_{i=1}^{n} \int_{t_0}^{t} a_i(t) \cdot p_i(t)\, dt + ManC$$

with $a_i(t)=a_i(t_0)(1+g_i(t))^t$ the amount of needed storage capacity in hierarchy level i at time t, and $p_i(t)=p_i(t_0)(1+d_i(t))^t$ the price for storage capacity in hierarchy level i at time t.

In chapters 3.1, 3.2 and 3.3 we have shown that there are useful ways to attach value to each single cost factor. In the next chapter we focus on the hardware cost and show that there are positive effects when employing ILM.

ANALYSIS OF ILM SYTSEMS

In ILM the storage hierarchies have different costs. Assuming hierarchy 1 is more expensive than hierarchy 2, etc., then there is a $\beta_i<1$ with $p_i(t)=\beta_i \cdot p_{i-1}(t)$ for all i≤n.

Further the amount of needed storage capacity $a_i(t)$ for hierarchy i is a fraction of $a(t)$: $a_i(t)=\alpha_i \cdot a(t)$ with $\alpha_i<1$ and $\sum_{i=1}^{n}\alpha_i = 1$.

β_i defines the price relation to the highest and most expensive hierarchy.

α_i defines the portion of the overall volume stored on hierarchy i.

We will show that α_i and β_i are effective parameters for deciding on an ILM solution. β_i is determined by the choice of employed technology. Choosing an appropriate technology is one task of ILM concept. α_i is determined by specific requirements of the related enterprise (ie. QoS requirements).

Assumptions

There are three assumptions to be made:

1. $d_i(t)$ are the same for all hierarchies.
2. $g_i(t)$ are the same for all hierarchies.
3. $d(t)$ and $g(t)$ are constant.

As shown in chapters 3.1 and 3.2 these assumptions are allowed and analysts give advice for their general conditions.

The assumptions simplify the model. With the simplified model the effects of price per capacity and amount of capacity needed are derived. The effects are still there when the assumptions are neglected. Therefore the assumptions are not necessarily needed to apply the results.

2-dimensional Cost

In order to get the ratio between 1-dimensional and 2-dimensional cost we divide $^{2}C(t)$ by $^{1}C(t)$:

$$\frac{^{2}C(t)}{^{1}C(t)} =: {}^{2}r(t)$$

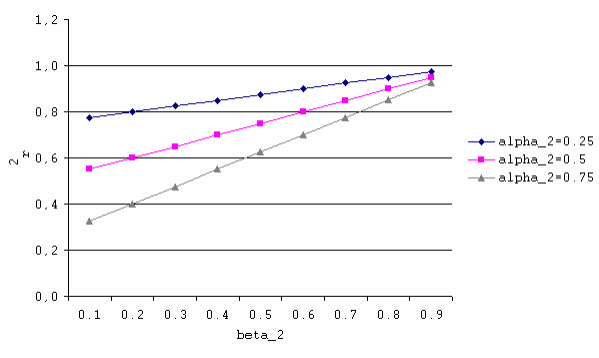
$$= \frac{\int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt + \int_{t_0}^{t} a_2(t) \cdot p_2(t)\, dt}{^{1}C(t)}$$

$$= \frac{\int_{t_0}^{t} \alpha_1 a_1(t) \cdot \beta_1 p_1(t)\, dt + \int_{t_0}^{t} \alpha_2 a_1(t) \cdot \beta_2 p_1(t)\, dt}{^{1}C(t)}$$

$$= \frac{\alpha_1\beta_1 \int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt + \alpha_2\beta_2 \int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt}{\int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt}$$

$$\sum \alpha_i=1,\beta_1=1,\beta_2<1 \quad \Rightarrow \quad {}^{2}r < 1 = \frac{^{1}C(t)}{^{1}C(t)} = {}^{1}r$$

Figure 2: Effects of α_2 and β_2 on ^{2}r



It is shown that ILM influences hardware cost. The effect shown with n=2 can be generalized for n>2. Each new hierarchy i, i>2, reduces the cost if there is an amount of data $a_i(t)=\alpha_i \cdot a(t)$ to be categorized in the hierarchy.

Effects of α_i and β_i on ^{2}r

To analyze how influential the employment of ILM is α_i and β_i have to be examined. In case of ^{2}r α_2 and β_2 have to be examined.

$$^{n}r = \sum_{i=1}^{n} \alpha_i\beta_i \cdot \int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt$$

To analyse the effects of α_i and β_i on ^{n}r the sum $\sum_{i=1}^{n}\alpha_i\beta_i$ is investigated.

For non-degenerated α_2 and β_2 Figure 2 shows the effects on ^{2}r.

It is shown that α_2 and β_2 each have impact on ^{2}r. If β_2 is close to 1 (=β_1) the effect is almost 0, in fact irrespective of α_2.

The effects on ^{2}r depend on the distance between α_i and α_2 and the distance between β_i and β_2.

Effects of α_i and β_i on ^{n}r

In the multidimensional case with n>2 similar effects like those for ^{2}rcan be expected. Since $\sum_{i=1}^{n}\alpha_i = 1$ there is a constraint influencing the distances between α_i.

We show two cases, where the effects of α_i and β_i are considered. Case 1 reflects a situation where the portion of hierarchy 1 is fix 50%. Although there is potential for all hierarchies n and the prices are well arranged, the influences of α_i and β_i cancel out each other.

Case 1:

n=2	$\alpha_1 = 1/2$	$\alpha_2 = 1/2$			
n=3	$\alpha_1 = 1/2$	$\alpha_2 = 1/4$	$\alpha_3 = 1/4$		
n=4	$\alpha_1 = 1/2$	$\alpha_2 = 1/6$	$\alpha_3 = 1/6$	$\alpha_4 = 1/6$	
n=5	$\alpha_1 = 1/2$	$\alpha_2 = 1/8$	$\alpha_3 = 1/8$	$\alpha_4 = 1/8$	$\alpha_5 = 1/8$

n=2	$\beta_1 = 1$	$\alpha_2 = 1/2$			
n=3	$\beta_1 = 1$	$\beta_2 = 2/3$	$\beta_3 = 1/3$		
n=4	$\beta_1 = 1$	$\beta_2 = 3/4$	$\beta_3 = 2/4$	$\beta_4 = 1/4$	
n=5	$\beta_1 = 1$	$\beta_2 = 4/5$	$\beta_3 = 3/5$	$\beta_4 = 2/5$	$\beta_5 = 1/5$

Now the relating ^{n}r for case 1 are considered.

$$^{n}r = \sum_{i=1}^{n} \alpha_i\beta_i \cdot \int_{t_0}^{t} a_1(t) \cdot p_1(t)\, dt .$$

Calculation of $\sum_{i=1}^{n}\alpha_i\beta_i$:

n=2: $\sum_{i=1}^{2}\alpha_i\beta_i$ =1/2·1+1/2·1/2=0.75

n=3: $\sum_{i=1}^{3}\alpha_i\beta_i$ =1/2·1+1/4·2/3+1/4·1/3=0.75

n=4: $\sum_{i=1}^{4}\alpha_i\beta_i$ =1/2·1+1/6·3/4+1/6·2/4+1/6·1/4=0.75

n=5: $\sum_{i=1}^{5}\alpha_i\beta_i$ =1/2·1+1/8·4/5+1/8·3/5+1/8·2/5+1/8·1/5=0.75

Case 2:

n=2	$\alpha_1 = 1/2$	$\alpha_2 = 1/2$			
n=3	$\alpha_1 = 1/3$	$\alpha_2 = 1/3$	$\alpha_3 = 1/3$		
n=4	$\alpha_1 = 1/4$	$\alpha_2 = 1/4$	$\alpha_3 = 1/4$	$\alpha_4 = 1/4$	
n=5	$\alpha_1 = 1/5$	$\alpha_2 = 1/5$	$\alpha_3 = 1/5$	$\alpha_4 = 1/5$	$\alpha_5 = 1/5$

n=2	$\beta_1 = 1$	$\beta_2 = 1/2$			
n=3	$\beta_1 = 1$	$\beta_2 = 2/3$	$\beta_3 = 1/3$		
n=4	$\beta_1 = 1$	$\beta_2 = 3/4$	$\beta_3 = 2/4$	$\beta_4 = 1/4$	
n=5	$\beta_1 = 1$	$\beta_2 = 4/5$	$\beta_3 = 3/5$	$\beta_4 = 2/5$	$\beta_5 = 1/5$

Now the relating ${}^{n}r$ for case 2 are considered.

$$ {}^{n}r = \sum_{i=1}^{n}\alpha_i\beta_i \cdot \int_{t_0}^{t} a_1(t)\cdot p_1(t)\,dt . $$

Calculation of $\sum_{i=1}^{n}\alpha_i\beta_i$:

n=2: $\sum_{i=1}^{2}\alpha_i\beta_i$ =1/2·1+1/2·1/2=0.75

n=3: $\sum_{i=1}^{3}\alpha_i\beta_i$ =1/3·1+1/3·2/3+1/3·1/3=0.667

n=4: $\sum_{i=1}^{4}\alpha_i\beta_i$ =1/4·1+1/4·3/4+1/4·2/4+1/4·1/4=0.625

n=5: $\sum_{i=1}^{5}\alpha_i\beta_i$ =1/5·1+1/5· 4/5+1/5·3/5+1/5·2/5+1/5·1/5=0.6

In case 2 each adding of a new information class creates a positive effect, of course. But it is shown that the advantage of adding a hierarchy becomes more and more marginal.

Summary Cases 1 and 2:

It is shown that α_i and β_i are the only individually adjustable parameters in an ILM scenario.

APPLICATION OF THE MULTIDIMENSIONAL COST MODEL

In chapter 4 we showed that α_i and β_i are effective factors for cost calculations and therefore for purchase decisions. The question is "How to determine these factors?". For the price factors β_i the answer is given by technology. The different storage technologies have different prices. The price relations are given by market analysts. The price difference between enterprise disk (FC, SCSI, FICON, ESCON) and midrange disk (SCSI, FC) is 2.5:1. The price difference between midrange disk and low cost disk (S-ATA) is 3:1. The price difference between low cost disk and automated tape is 6:1 [7].

These are quite useful results and complete the estimations given in chapters 3.1, 3.2, 3.3. What is missing are the potential factors α_i. How many gigabytes are needed in the different hierarchies? Is there enough potential for ILM? To get these figures we conducted a case study in spring 2005. In the case study we investigated a database of a german DAX-30 company. The access patterns of 150,000 files were investigated. A random sample of 1,000 files was taken and all accesses were logged. The logging of all accesses since creation of each file provided the following results [12]:

There were more than 150,000 files on the system and 89 percent of them were not accessed 90 days after their creation.

Therefore in this case the potential factors are:

$$ \alpha_1 = 0.11 \text{ and } \sum_{i\neq 1}^{n}\alpha_i = 0.89 . $$

The question is "How many hierarchies?". Since currently there is one hierarchy only, the answer is 2 or 3.

Assuming there are requirements for more 3 hierarchies the potential of 0.89 has to be distributed over hierarchies 2 and 3. Let be α_2=0.3 and α_3=0.59 (α_1=0.11). Then a 3-dimensional-ILM solution of a 1.3 TeraByte database would, for example, look like:

Total volume: 1.3 TeraBytes

Storage hierarchy 1:

Enterprise Disk (FC), required capacity: 143 GB

Storage hierarchy 2:

Low Cost Disk (S-ATA), required capacity: 390 GB

Storage hierarchy 3:

Automated Tape, required capacity: 767 GB

The price relation is:

FC : S-ATA : Automated Tape =

1: 1/7.5: 1/45 = β_1:β_2:β_3

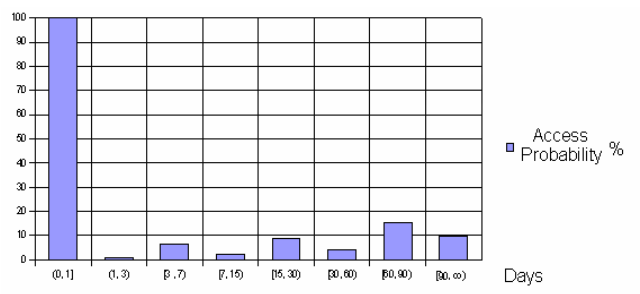
By determination of these key figures the ILM cost effects are characterised sufficiently, and IT-managers are supported to find an informed cost decision.

CONCLUSION AND OUTLOOK

Based on cost considerations we have shown that decisions about ILM solutions depend on enterprise external and enterprise internal factors. External factors like market prices for hardware are the same for all ILM solutions. The internal factors can be influenced individually by the enterprise and need to be taken into account separately. The conclusions are:

1. (n-dimensional) ILM has proven positive effects on cost.
2. The change from no ILM to 2-dimensional ILM has the biggest

Figure 2. Access probabilities



gain.

3. The change 2-dimensional ILM to n-dimensional ILM does not guarantee gains.
4. For each new ILM-dimension the potential α_i and the price factor β_i have to be considered.

If we sum up the detailed conclusions, ILM systems can offer significant cost savings for enterprises if they are used correctly.

Our next step will be the simulation of actual ILM systems in order to get reliable statements for migrating information. Furthermore the case study will be extended to derive distribution functions for file accesses.

REFERENCES

[1] G. Moore, Cramming more components onto integrated circuits, Electronics, volume 38, number 8, 1965.

[2] A. Lyman, How Much Information? 2003, University of California, October 2003

[3] D. Fletcher, Benchmark and Trend Analysis, Division of Information Technology Services, State of Utah, October 2003

[4] H. Nguyen, IDC's Worldwide Disk Storage Systems Quarterly Tracker, March 2005

[5] F. Moore, Storage - New Game New Rules, Horison Information Strategies, 2003

[6] T. Kraemer, The Storage Report - The customer Perspectives & Industry Evolution, McKinsey & Company, June 2001

[7] R. Paquet, Why You Need a Storage Department, Gartner Research, June 2004

[8] M. Satyanarayanan, A Study of File sizes and Functional Lifetimes, Proceedings of the 8th Symposium on Operating Systems Principles, Association of Computing Machinery, 1981

[9] S. Strange, Analysis of Long-term Unix File Access Patterns for Application to Automatic File Migration Strategies, University of Berkeley, 1992

[10] L. Turczyk et al. Analyse von Datei-Zugriffen zur Potentialermittlung fuer ILM, TU Darmstadt KOM Technical Report 01/2005