CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF INFORMATION TECHNOLOGY

ASSIGNMENT OF MASTER’S THESIS

Title:	Analysis of Backup for Small and Medium-sized Enterprises (SME) in the Czech Republic
Student:	Bc. Jaroslav Vašák
Supervisor:	Dr. Pi Chung Wang
Study Programme:	Informatics
Study Branch:	Computer Systems and Networks
Department:	Department of Computer Systems
Validity:	Until the end of summer semester 2017/18

Instructions

1. Define the term data backup and associated terms.
2. Design a model for a Cost-Benefit Analysis of hybrid cloud backup solutions.
a. Define the cost and benefit factors of backup solutions.
b. Design a cost-benefit model.
c. Describe how individual components of the model are calculated.
3. Create a tool for calculating this model.
4. Utilize this model to different types of businesses within SMEs using different types of backup solutions and use the tool to evaluate the precise numbers.
5. Based on the previous results, create a set of general recommendations for the backup in SMEs applying both technological and managerial dimension terms.
6. Perform the research of existing tools and resources within the Czech Republic that allow performing a backup of data to a cloud environment.
7. Within your resources, test the tools to verify your previous recommendations.
8. Create a new set of recommendations applicable only for the Czech Republic.

References

Will be provided by the supervisor.

prof. Ing. Róbert Lórencz, CSc.
Head of Department

prof. Ing. Pavel Tvrdík, CSc.
Dean

Prague February 16, 2017

Czech Technical University in Prague

Faculty of Information Technology

Department of Computer Systems and Networks

Master's thesis

Analysis of Backup for Small and Medium-sized Enterprises (SME) in the Czech Republic

Bc. Jaroslav Vašák

Supervisor: Dr. Pi Chung Wang

29th June 2017

Acknowledgements

The last year was a great challenge for me. Not only because I was working on my thesis the whole time, but for eight months I studied in Taiwan, a country with a different culture, customs and mainly a language barrier.

I would like to thank all those who have been supporting me all the time and helped me to complete this thesis without them this project would not have been possible.

I am especially indebted to Ing. Pavel Náplava who helped me find an interesting topic and helped me all the time in general, despite all the distance that separated us till the end. I am deeply grateful to Dr. Pi Chung Wang, the person who led me at the foreign university in Taiwan as my supervisor and helped me a lot of times to keep moving in the right direction with his insightful comments and suggestions.

A special thanks to my girlfriend Rita Chang for her love, support and also listening to me even though I was not always full of optimism and enthusiasm.

My deepest appreciation goes to Monika Lajdová who offered me her professional language proofing at the end of my thesis. Furthermore, I would like to thank my parents, especially to my mom, who created a great environment at home after my arrival.

I also thank all my friends who have had a great deal for understanding me and have been there for me whenever I needed to talk. I am grateful to all of those I met in Taiwan. Above all, I appreciate my awesome roommates Olly, Jelle and Ruben, who kept my optimism even during the worst of times. Finally, I would like to thank my dearest friends who were waiting for my return at home, namely Martin, Hai, Markéta and Michaela. You guys are the best!

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Article 46(6) of the Act, I hereby grant a nonexclusive authorization (license) to utilize this thesis, including any and all computer programs incorporated therein or attached thereto and all corresponding documentation (hereinafter collectively referred to as the “Work”), to any and all persons that wish to utilize the Work. Such persons are entitled to use the Work for non-profit purposes only, in any way that does not detract from its value. This authorization is not limited in terms of time, location and quantity.

In Prague on 29th June 2017 . . . .

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis

Vašák, Jaroslav. *Analysis of Backup for Small and Medium-sized Enterprises (SME) in the Czech Republic.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2017.

Abstrakt

Cílem této práce je navrhnout malým a středním podnikům adekvátní typ nasazení zálohovacího řešení. Zálohování dat je i přes jeho důležitost stále mnohými firmami opomíjeno.

K řešení problému je navrhnut matematický model, který vychází z analýzy nákladů a přínosů zálohovacích řešení. Cíl práce je naplněn vytvořením podpůrného nástroje, sloužícího k získání konkrétních hodnot z modelu, které jsou posléze vyhodnoceny.

Na základě vyhodnocení nezávislé analýzy je vytvořen soubor jak obecných, tak i konkrétních doporučení, kam zálohovat v rámci České republiky na základě celkové velikosti zálohovaných dat.

Výsledky práce umožňují pomoci malým a středním podnikům při rozhodování, zdali využít lokálního, cloudového, anebo hybridního zálohovacího řešení, na základě ceny, anebo přínosů.

Klíčová slova zálohování dat, malé a střední podniky, celkové náklady na vlastnictví, analýza nákladů a přínosů, matematický model, hybridní cloudové zálohování, AWS, Azure, Google

Abstract

The aim of this thesis is to design an adequate type of backup solution deployment for small and medium-sized enterprises. Data backup, despite its importance, is still neglected by many companies.

To solve this ignorance, a mathematical model based on a cost-benefit analysis of backup solutions is proposed. The objective of the thesis is fulfilled by the creation of a supporting tool for obtaining specific values from the model, which are then evaluated.

Based on the evaluation of the independent analysis, not only general and but also specific recommendations related to the storage location of backup within the Czech Republic is created based on the total size of the backed up data.

The results from the thesis enable small and medium-sized businesses to decide whether to use a local, cloud, or hybrid backup based on the cost of the solution and its benefits.

Keywords data backup, small and business-enterprises, SME, total cost of ownership, TCO, cost-benefit analysis, mathematical model, hybrid cloud backup, AWS, Azure, Google

Contents

List of Figures

List of Tables

Introduction

The times when cloud computing was just a mere buzzword is over. People who are still afraid of cloud security today are slowly declining and it is time to ask other questions. Does the cloud solution really offer the benefits it presents? And if so, how is it possible that we all are still not using it?

In this thesis, I will perform an independent analysis of backup solutions for small and medium-sized companies, especially in the Czech Republic. The thesis focuses on small and medium-sized companies because of their numerical superiority. These companies account for 99% of all businesses, and therefore they are more likely to decide which solutions to choose from. The decision of these firms complicates the lack of independent studies and the lack of a number of independent tools because they often implement an inappropriate solution.

During this thesis, we proceed from the following hypotheses. We assume that cloud storage is cost-effective and offers the best price for 1 GB. Its disadvantage is the need to use a slower wide area network, which leads to slower data backup and recovery. That is why we assume that a hybrid backup solution is the best way to backup data, it combines the advantages of a cheap cloud storage with a fast local backup.

Aims and objectives

The aim of the thesis is to make an independent study of the current backup options for small medium-sized businesses. Based on which, we will present a set of general and specific recommendations for the Czech Republic. This answers the question in which cases is it preferred to backup locally, into the cloud, or to use hybrid both in terms of cost and benefit from individual backup solutions.

To achieve this aim, I will:

1. Define a cost and benefit factors of backup solutions.

2. Design a mathematical model for a cost-benefit analysis of hybrid cloud backup solutions.

3. Create a data backup comparison tool for calculating designed model.

4. Analyse different types of backup solutions in several small and medium-sized model businesses divided by the size of the backup data.

5. Perform a research of existing backup products within the Czech Republic.

6. Test the found backup products.

Literature review

With the aim of finding the literature engaged in the backup analysis for SMEs, were used a combination of keywords in research databases such as ACM Digital Library, IEEE Explore, SemanticScholar, Web of Science, SpringerLink and Google Scholar. The combination of keywords consisted of data backup, cloud computing, cost model, total cost of ownership, in-house, SME, cost-benefit, analysis, comparison and hybrid.

It was founded several papers dealing with the similar themes. The paper [1], [2], [3], [4] discuss cost factors in context of cloud computing, thereof the papers [1] and [2] present cost model and the paper [2], [3] and [4] contains general comparison between private and public cloud. The paper [5] discuss cost factors in the context of cloud storage and the paper [6] evaluate disaster recovery plan (DRP) from the perspective of different backup sites and it presents basic cost model.

A detail cost model was demonstrated in [1], with the focus on large enterprises with own in-house data center and the option to migrate some of the tasks to the Cloud. The cost model is designed for hybrid cloud to determine which services is better to migrate from private cloud to public cloud to save our expenses.

A three-layer cost-benefit model introduced in [2], helps us to decide when and in what situation is recommended to shift to cloud computing, or stick with in-house solutions. The first layer is used to base cost estimation. The second layer is used to analyze the data pattern based cost and the third layer is used for a specific project.

A case study of SME startup was presented in [3], the TCO were compared with Amazon Web Services (AWS) cloud with the help of tool is not available.

The question for schools whether to shift to cloud or not were discussed in [4]. The cost benefit analysis of the TCO and cloud were performed for 30 users, which falls under the typical SME.

Application of cost model for data center was performed in [5]. The cost of local small datacenter was compared with Amazon S3 to determine the full cost of the cloud based storage system.

A basic cost model of disaster recovery solution is introduced in [6].

However, no research investigating the backup in a hybrid cloud from the perspective of the cost-benefit analysis of SMEs was found. This may be due to a simplicity of the problem.

There are several public tools available on the Internet that focus on TCO of cloud and on-premises or on data storage. One of them is a tool from SherWeb [7] that comparing the total cost of ownership of on-premises virtualization farm with cloud-based infrastructure as a service (IaaS). Another is a tool from SNIA [8], which comparing TCO of HDD and SSD. Cloud providers such as Google, AWS, and Azure also offer public TCO calculators [9], [10] and [11].

However, no public tool comparing the total cost of ownership of cloud, on-premises and hybrid, along with the benefits of each solution has been found.

Definition of terms

Availability - probability that the system is available during certain time periods when its use is required, usually the value is given in a percentage within a year [12].

Data archiving - is the practice to identify and move inactive data out of the currently used production systems into the long-term archival storage systems. Shifting unused data from production systems enhancing the performance of resources needed while archival systems store data more cost-effectively and provide data for retrieval when necessary [13].

Data backup - is a copy or duplicate version of a file, program, or entire volume, held for use if the original data is in some way damaged [14].

Data compression - is the way how to reduce the data and consume less space on disk by modifying, encoding or converting the bits. It is also known as source coding or bit-rate reduction [15].

Data deduplication - is a process of reducing the data by removing the duplicate copies from the storage. During the removal of extra copies just one copy of the data is saved and the others are replaced with the pointers, pointing to the original copy [16].

Durability - the ability of equipment, media, or machine to exist for a long period of time without significant degradation by resisting the effects of heavy use, wetting, heating, freezing, corrosion and oxidation [17].

Data durability - durability with respect to data is the probability that data remains unaffected and accessible after one year. In other words, the probability of data not lost [18].

Recovery Point Objective (RPO) - is limited by maximum amount of data loss tolerated and is measured in time. It is a time of data in a backup storage required to regain normal operations if some accident happens [19].

Recovery Time Objective (RTO) - is the maximum required period permitted between sudden failure or disaster and the restoration of normal operations and service levels. It defines length of time after a failure or disaster when the consequences of the interruption turn unacceptable [20].

Small and medium-sized enterprises (SMEs) - are the enterprises having less than 250 persons employed with the annual turnover up to EUR 50 million, or a balance sheet total of no more than EUR 43 million [21].

Chapter 1

Data Backup

In this chapter, a theoretical background with the focus on data backup techniques and strategies will be presented. This is mainly important for people who do not have a complete knowledge about this thesis topic. It can also serve to the others as a reminder of important backup factors or as a summary of a recent information related to data backup and data archiving.

First, the types of backup will be introduced. Second, the important backup techniques which are used to increase efficiency of data backup will be examined. Third, the possibility to apply backup to the cloud, followed by the basic storage media used for data backup and data archiving along with the RAID technology will be presented. Then the individual media in terms of advantages and disadvantages will be compared and from that comparison an appropriate architecture for backup will be presented. Furthermore, the calculation of storage availability will be unveiled, because of the comparison of local and cloud availability. Finally, a general recommendation for data backup will be introduced.

1.1 Backup types

There are many ways of categorizing the different types of backup. Dilip C. Naik divided the backup into three categories: Architecture, Functionality, Network Infrastructure [22]. Umesh Hodeghatta Rao divided the backup into four categories: based on current data on the system and the data on the backups, based on what goes into the backup, based on storage of backup, based on the extend of the automation of the backups [23].

In the following subsections, we will divide the types of backup into four categories: based on functionality, based on storage location, based on architecture and based on deployment model.

Category 1: Based on functionality

- Full Backup: includes all the data in our system selected to be backed up, regardless of when they were last modified. It is required for all other types of backup [24].

- Incremental Backup: includes only those files that have changed since the last backup whether it was a full or an incremental backup. The restore operation involves the last full backup plus all the following incremental backups [25].

- Differential Backup: includes only those files that have changed since the last full backup. In case of restoration, it involves the full backup plus the last differential backup [25].

- Mirror backup: is a reflection of the source being backed up. A real time duplicate. Some do not consider a mirror to be a backup, because with the mirror backups, when a file in the source is deleted, that file is eventually also deleted in the mirror backup. For that reason, the mirror backups should be applied with awareness of accidental file deletion, sabotage or virus invasion. Data compression and password protection cannot be applied to the files in the mirror backup [26].

- Synthetic backup: is a process of creating full backup from the previous full backup and subsequent incremental backups. It can be used in case of limited bandwidth. The procedure is called ”synthetic” since it is not a backup created from original data [27].

The list above includes basic types of backups, such as full backup, incremental backup and differential backup. The list is not complete, because today’s backup products usually offer several combinations of basic types of backups and techniques already mentioned such as synthetic backup. For example, Veeam offers these backup methods: forever forward incremental backup, forward incremental backup and reverse incremental backup (for more information see the white paper called Veeam backup methods and the impact on destination storage I/O from Lucca Dell’Oca).

The following table 1.1 provides an overview of backup types based on functionality.

Category 2: Based on storage location

There are two ways of backup based on storage location.

- On-site backup: is any backup where the storage medium is stored locally. The storage medium could be connected directly, or through a local area network to the source being backed up [28].

Table 1.1: Comparison of backup types based on functionality.

Backup type	Data backed up	Backup time	Restore time	Storage space
Full backup	All data	Slowest	Fast	High
Incremental backup	New or modified	Fast	Moderate	Lowest
Differential backup	New or modified since last full	Moderate	Fast	Moderate
Mirror backup	New or modified	Fastest	Fastest	Highest

- Off-site backup: is a backup where the copies are kept in a different geographical location such as another city or cloud storage [29].

Category 3: Based on architecture

There are two approaches to performing a backup.

- File-level backup: consists of specific files from the environment with the possibility to not include some high-level file system data such as file permissions. There are two cases presented below:
  - Application-aware backup: operates on application level and hides the file structure underneath. For instance, a database-aware backup can include database file and transactions logs as one unit.
  - System State backup: is an option for Windows OS providing a default set of operating-system-specific files that are vital for OS [30].
- Image-level backup: this type of backup is mostly called image-level backup or block-level backup, but could be also named as bare metal backup/recovery (BMR), disaster recovery backup, volume-level backup, ghost backup, or clone of your machine [31]. The image-level backup is saved as a single file/image that contains a copy of the operation system with all connecting data, including the system state and application configuration at a certain time. It can be used for a computer or virtual machine (VM)[32].

The following table 1.2 provides an overview of backup types based on architecture.

Table 1.2: Comparison of backup types based on architecture.

Backup type	Advantages	Disadvantages
File-level backup	• Efficient for restoring a small number of files • Flexibility - backup and restoration of individual files and folders on a volume • Flexible backup policies for different data types on the same volume	• Time consuming - when too large or to many small files are backed up • A small change in a large file means, the entire file must be backed up again
Image-level backup	• Efficient for restoring the whole system • Fast restore time • Fewer performance issues	• Not efficient for restoring a small number of files • Requires more storage space

Category 4: Based on deployment model

- Local backup: data are backed up to a locally available storage (see on-site backup 1.1)
- Cloud backup (online backup): is sending backup data directly to the cloud through proprietary or public network. Data are stored off-site into a provider's cloud or to a public Cloud Service Provider (CSP) like AWS, Azure or Google [33].
- Hybrid cloud backup: is a mix of local backup and cloud backup. It consists of an on-premises appliance that should store at least one full backup along with the following incremental backups. The data are first stored locally and then replicated to the CSP [34].

The following table 1.3 provides an overview of backup types mentioned above. I intentionally omitted many times repeated advantage of public clouds and that it is more cost efficient/cheaper. Whether is a public cloud cheaper or not we will see in the next chapter 2 *Cost-Benefit Analysis and design of hybrid cloud backup solutions.*

1.2 Backup techniques

In this subsection will be introduced techniques that are used for a data backup to achieve a better cost effective solution. These techniques arose due to exponential data growth [35, p. 73] and [36]. Techniques such as data compression

Table 1.3: Comparison of backup types based on deployment model.

Backup type	Advantages	Disadvantages
Local backup	• Fast backup and restore	• Complexity • Potential single location problem (disaster)
Cloud backup	• Durability • Scalability • High availability • Accessibility • No single location problem (in case of multi-regional solution)	• Dependence on the internet connection • Backup and restore is limited by bandwidth • Jurisdiction problem (different location of sensitive data)
Hybrid backup	• Fast backup and restore with proper backup arrangement • Improved durability • Improved availability • Enhanced redundancy (local and cloud backup)	• Complexity

and data deduplication help reduce the size of data in on-site storage or in off-site storage.

With the growing trend of cloud service usage, the need of internet with sufficient bandwidth increased as well. Because of that in the backup we use a group of techniques that are expanding the efficiency of data transfer and reducing the network traffic. This set of techniques is called wide-area network (WAN) optimization.

Data deduplication

Data deduplication is a technique for reducing storage capacity and is ideal for highly redundant operation like backup. In the following subsections, we will go through options which are offered by today's tools. We will describe their advantages, disadvantages and usage.

Chunking methods

Data deduplication can be done at variable levels of granularity.

- File-level: deduplication takes place on the file level. The entire file is considered as a chunk. Only one index is created for one file.

- Block-level: deduplication looks within a file. The file is broken into the chunks.
    - Fixed-size: the file is split into proportionally equal chunks. The size of the chunk could be usually chosen.
    - Variable-size: the file is separated into multiple chunks of variable sizes based on the content of the file. One of the most common algorithm of variable size chunking is Rabin's algorithm [37].

Using a more complex algorithm leads to a greater storage saving but in the same time it leads to an increased system load and elongated processing time. On the other hand, simpler file-level deduplication need less system resources and takes less processing time. The deduplication ratio is usually not that good, because a small modification in a large file resulting in the need to make a new copy of the file. In [38] Meyer and Bolosky have found, that file-level deduplication achieves about 75% of the space saving in the Rabin's algorithm for storing the live file system and 87% for the backup images. That is not a big difference compared to how much system resources we can save with the easier file-level deduplication method. However, condemning other more complex methods could be a mistake, so we should take more complex methods into consideration during our next backup planning.

The following table 1.4 provides a summary of chunking methods.

Table 1.4: Basic comparison of chunking methods.

Chunking method	**Storage saving (deduplication ratio)**	**System load and backup delay**
File-level	Lowest	Lowest
Fixed-size	Moderate	Moderate
Variable-size	Highest	Highest

Deployment type

In general, deduplication can be deployed in one of two ways:

- At source: deduplication at the source removes redundant blocks before the data are transmitted to the backup system. This saves network bandwidth and accelerates transfer to the cloud. More resources of source node are needed.

- At target: deduplication at the target takes place within the backup system after the data are transmitted. More bandwidth than at source is needed and system resources are utilized especially at the target. There are two ways on how to deduplication at the target:

- In-line data deduplication: takes place before deduplicated data is written.
- Post-process data deduplication: takes place after the data is stored [12].

Furthermore, we can deploy a global data deduplication, which could be a solution suitable for larger companies. Here is a generic definition from TechTarget [39]:

> *"Global data deduplication is a method of preventing redundant data when backing up data to multiple deduplication devices. This situation may involve backing up to more than one target deduplication appliance or, in the case of source deduplication, backing up to multiple backup nodes that are."*

Basically, if our data are sent across WAN we should consider deployment at source, which will reduce the amount of data sent and speed up the overall data transfer. On the other hand, if we are using deployment at source then we will have multiple virtual machines sharing one physical host. Running multiple hash calculations at the same time without enough free resources at the physical host may overload the host server.

The following table 1.5 provides a summary of at source and at target deployment.

Table 1.5: Basic comparison of deduplication deployment type.

Deployment type	**At source**	**At target**
Endpoint device resources	High	Low
Backup storage resources	Low	High
Bandwidth savings	High	None

In-line deduplication reducing the disk capacity needed, but processing the deduplication could cause a bottleneck that can affect the length of the backup time. Conversely post-process deduplication minimizes the backup time, there is no backup delay, but it needs more storage capacity because the data are first saved undeduplicated and then deduplicated.

The following table 1.6 provides a summary of in-line and post-process data deduplication techniques.

Deduplication ratio

Deduplication ratio is the ratio of data's original size to data's size after deduplication. This ratio is used to estimate how much space do we save with dedupe and if it is worth it to deduplicate our data.

Table 1.6: Basic comparison of at target deduplication deployment options.

Deduplication techniques	**In-line**	**Post-process**
Backup speed	Delayed by processing	Normal
Storage capacity needed	Low	High

A simple equation 1.1 using numbers of *unique data percentage* is derived from Acronis experience with its customers used for empty storage to estimate a deduplication ratio can be found in a white paper [40] and may vary in your environment.

$$Deduplication\ ratio = \frac{Unique\ data\ percentage}{100\%} + \frac{1 - \frac{Unique\ data\ percentage}{100\%}}{Number\ of\ machines} \tag{1.1}$$

Unique data percentage:

- Virtual machines: 30 percent unique
- Office workstation: 50 percent unique
- Database servers: 65 percent unique
- File servers: 75 percent unique

The deduplication ratio in the backup is growing with the time and amount of copies stored. The time is given by a retention period and the amount of copies are affected mostly by the full backup, so it depends on our backup scheme.

WAN optimization

WAN optimization is a series of technologies and techniques for maximizing the efficiency of data flow across the WAN. Techniques like data compression and data deduplication were mentioned in the previous subsection 1.2, some of the other WAN optimization technologies are listed below [41]:

- Traffic shaping - is prioritizing certain packets inside the data flow so the bandwidth is distributed accordingly.
- Data caching - the regularly accessed data are stored locally for faster access.
- Protocol spoofing - is optimization of slow and chatty protocols.
- Latency optimization.
- Forward error correction - is a technique for reducing the retransmission of packets.

Summary

Techniques mentioned above brings us advantages such as reducing storage space/network cost or accelerating access to off-site storage. However, there are also drawbacks e.g. when latency emerges out of data compression or data deduplication with the possibility of RTO increasing. Next, we usually need to buy additional HW and SW for certain technology use. Furthermore, data deduplication has impact on the performance of local or remote node. Finally, implementing WAN optimization could be costly and more complex.

1.3 Cloud backup storage

In this chapter, the possibility to apply backup to the cloud for the three leading CSPs (AWS, Azure and Google [42]) along with a variety of cloud storage suitable for backup or archiving will be unveiled. That is because these CSPs are offering a range of options from small businesses to large enterprises. On the market are lots of other online backup services such as Backblaze B2 Cloud Storage, CrashPlan PRO, OpenDrive for Business, Carbonite Business and more. These smaller CSPs will not be included in the overview, because their products are often targeted at a specific market and they do not offer such variability of services like the largest CSPs.

Storage Classes

The leading CSPs offers public cloud storage classes based on the retrieval frequency. To compare different cloud storage classes, we grouped all storage services into four tiers:

- Hot: durable, highest availability, geo-redundant and appropriate for storing data that is frequently accessed, such as serving website content, interactive workload, or data supporting mobile and gaming applications. Representatives are:
  - Amazon S3 Standard + Cross-Region Replication (CRR)
  - Google Cloud Storage (GCS) Multi-Regional
  - Microsoft Azure Hot Blob Read Access - Geo Redundant Storage (RA-GRS)
  - Microsoft Azure Hot Blob Geo Redundant Storage (GRS)
- Warm: highly available, lower cost, less durability and is determined for frequently accessed non-critical reproducible data. Representatives are:
  - Amazon S3 Standard
  - Google Cloud Storage Regional

  - Microsoft Azure Hot Blob Zone Redundant Storage (ZRS)
  - Microsoft Azure Hot Blob Locally Redundant Storage (LRS)

- Near-line: low-cost storage class for storing data that is accessed less frequently, but requires rapid access when needed, e.g. backup. Representatives are:
  - Amazon S3 Standard - Infrequent Access (IA)
  - Google Cloud Storage Nearline
  - Microsoft Azure Cool Blob GRS
  - Microsoft Azure Cool Blob ZRS
  - Microsoft Azure Cool Blob LRS
- Cold: very-low-cost storage, highly durable storage service for data archiving, online backup, and disaster recovery. Representatives are:
  - Amazon Glacier
  - Google Cloud Storage Coldline

The following table 1.7 provides an overview of prices and the availability of storage classes offered by leading CSPs. In the overview, other charges associated with using cloud storage such as cost per operation, or cost per data retrieval are not shown. There are also limits not mentioned (e.g. minimum storage duration). The table is actual to date April 2017 and the information was taken from the official sites of CSPs [43], [44], [45], [46], [47], [48] and [49].

The data from the previous table are shown in graphical form in the following figure 1.1.

1.4 Backup storage architecture

In this section, the storage media used for data backup and data archiving will be introduced. Then the RAID technology used by the media along with a problem may arise in a certain RAID as well as the recommendation of which RAID to use will be presented. Furthermore, both general and specific terms of the storage media will be compared. Finally, based on the previous comparison, the data backup architecture will be presented. This information can be used for planning local backup or archiving solution.

Data backup storage media

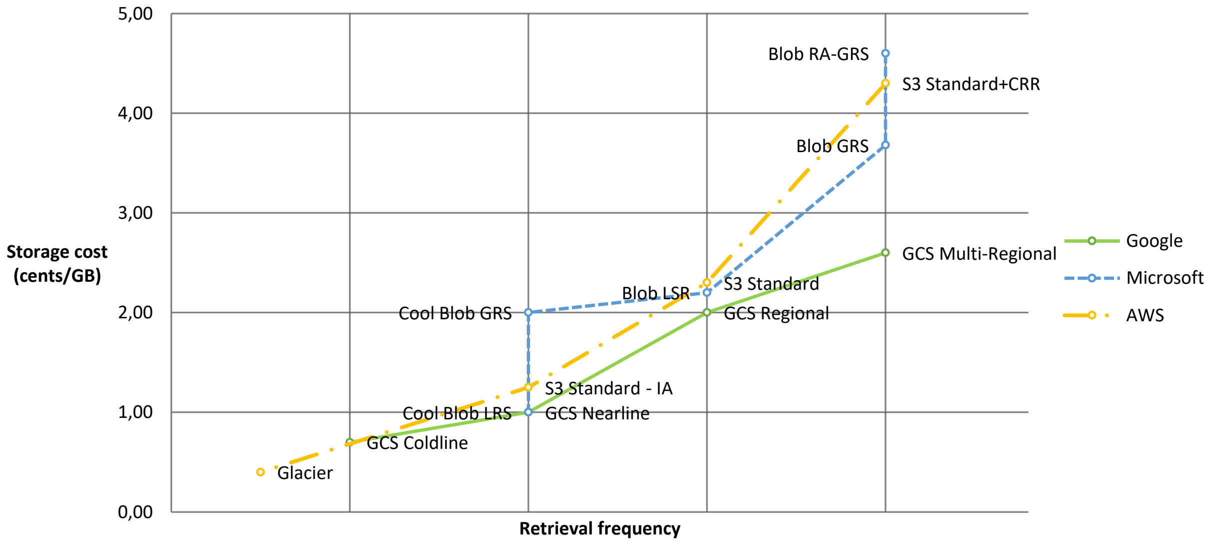
The data from backups can be stored on one or more of the following storage media:

Table 1.7: Comparison of public cloud storage classes.

Storage class		Availability[1]	Monthly price per GB[2]
Hot	Hot Blob RA-GRS	99.99%	$0.046
	S3 Standard + CRR	99.9%	$0.043[3]
	Hot Blob GRS	99.9%	$0.0368
	GCS Multi-Regional	99.95%	$0.026
Warm	Blob ZSR	99.9%	$0.0296
	S3 Standard		$0.023
	Hot Blob LSR		$0.022
	GCS Regional		$0.02
Near-line	Cool Blob GRS	99%	$0.02
	S3 Standard - IA		$0.0125
	Cool Blob LRS		$0.01
	GCS Nearline		$0.01
Cold	GCS Coldline	99%	$0.007
	Glacier	N/A	$0.004

[1] Availability defined in the Service Level Agreement (SLA)
[2] The storage cost per GB for first 50 TB of data stored.
[3] Plus additional charges for data transfer between AWS regions.

Figure 1.1: Comparison of cloud storage classes.



- Solid-state storage: is typical storage using non-volatile memory, it retains data when power is shut off. It stores and retrieves data by using integrated circuits without any involvement of moving mechanical parts. Among the solid-state storages belongs the following devices [50]:
  - Solid-state drive (SSD)
  - Secure Digital (SD)
  - USB flash drive
- Hard disk drive (HDD): is a data storage device that uses magnetic storage to store and retrieve data using one or more rapidly rigid rotation platters.
- Magnetic tape: is a storage medium with sequential access to data. The tape technologies were categorized as entry level, midrange and enterprise. Midrange tape generally meant LTO technology which has been the predominant tape technology in the market by the volume shipped. By IDC the enterprise tape technology belongs to proprietary tape drive like the TS11x0 and T10000. Today's newest and most widely used tapes are: LTO-7 and TS1150. According to the SPECTRA's white paper [51], users with less than 1 PB of data and low data growth expected should use LTO-7, conversely for users with 1 PT or more, use of TS1150 technology is superior.
- Optical storage: is an optically readable medium such as CD, DVD, Blu-ray and Optical Disc Archive. Optical Disc Archive is a system developed by Sony with the goal to make durable, scalable and long-term archiving solution mainly for datacenters with a product called Everspan [52]. Standard as M-disk, the storage medium with the longest life span theoretically is 1000 years.

RAID levels

A redundant array of independent disks (RAID) is used to increase the storage reliability, capacity and performance.

- RAID level 0: the data are evenly split up across two or more disks, that is called striping. This brings better Read/Write (R/W) performance, but also a higher chance of disk failure, therefore it is not used alone.
- RAID level 1: the exact copy of data is stored on two or more disks, that is called mirroring. It enhances the reliability and read performance with every disk added.
- RAID level 5: data blocks are striped across the drivers with distributed parity. The minimum configuration needed are three disks. Read

operations are enhanced, while write operation are degraded and have one fault tolerant disk. In case of failure of one disk the parity is used for recalculating the missing data and results in slower read operations.

- RAID level 6: extends RAID 5 by adding another parity block. This brings us another fault tolerant disk, so the system is functional in case of two disk failures. The write performance adds little bits of overhead than the RAID 5 case, but the read performance is almost identical. The minimum configuration needed are four disks.
- RAID level 10: is combination of RAID 1 and RAID 0, the data are first mirrored across two disks then striped across each set of drives. It combines the benefit of a faster R/W with higher reliability. The minimum configuration needed are four disks.

Problem of reconstruction RAID level with parity

Adam Leventhal's paper [53], mentioned the need to add more parity to maintain the reliability. Robin Harris continued in Leventhal's paper with his article [54], he mentioned that RAID 5 is not a recommended solution anymore and the same situation could even be for RAID 6 in the future.

Articles described that in case of RAID levels using parity (RAID 5 and RAID 6) we are faced with a problem in reconstruction of the array after a disk failure. This problem is caused by unrecoverable read error (URE) per bits read, its value can be found in the specification, e.g. for consumers HDD is typically URE 1 *in* 10^{14} and for the enterprises HHD is 1 *in* 10^{15}.

When a disk in RAID level with parity fails and is replaced, all the data on the other drivers in the array are read to reconstruct the array to its original state. The more data are read the higher chance we encounter the URE. If the URE happens during the RAID 5 rebuild, one or more data blocks are lost and in a worst case, we lose everything due to array failing to rebuild.

Real example

How to calculate the probability of failure during the RAID with parity rebuild is described in IBM's blog article [55]. The equation to get probability of URE is:

$$100 * (1 - (1 - e)^b) \tag{1.2}$$

Where:

- e = the non-recoverable read error per bits read (error rate)
- b = number of bits read (size of the volume in bits)

For the calculation, two currently available HDD designed for backup/archive were taken from the market (March 2017). HDDs specifications:

- disk A - capacity 8 TB, URE = 1 *in* 10^{14}, AFR = 1.095%, data transfer = 150 MB/s
- disk B - capacity 12 TB, URE = 1 *in* 10^{15}, AFR = 0.35%, data transfer = 261 MB/s

The following table 1.8 shows a percentage of RAID 5 and RAID 6 failure during the rebuild. As we can see RAID 5 is highly unreliable. RAID 6 is greatly reducing the probability of rebuild failure, but even like that the probability of failure especially with more high capacity disks is not negligible.

Table 1.8: Rebuild failure of RAID with parity caused by URE.

RAID composed of		5 x disk A	10 x disk A	5 x disk B	10 x disk B
RAID 5	Capacity	32 TB	72 TB	48 TB	108 TB
	Drive failure	0.029	0.15	0.008	0.04
	Probability of URE	92.27	99.68	31.86	57.82
	Rebuild failure	**92.299**	**99.83**	**31.868**	**57.86**
RAID 6	Capacity	24 TB	64 TB	36 TB	96 TB
	Drive failure	0.016	0.12	0.005	0.03
	Probability of URE	85.32	99.4	25	53.58
	Rebuild failure	**1.37**	**11.93**	**0.125**	**1.6**

SSD RAID

The behavior of SSD in RAID configuration is not certain yet. Previously we assumed that SSD reliability depends on the number of write cycles. In RAID with parity we have twice more write operating, due to writing parity, so the reliability is smaller [56].

But the paper [57] brings new key conclusions. First the SSD age, not usage, affects reliability. Second, ignore the uncorrectable bit error rate (UBER) specs. And finally, SSDs fail at a lower rate than disks, but UBER rate is higher. Which would mean an even less probability to rebuild RAID configuration with parity.

Summary

In the following table 1.9, is a summary of RAID 6 and RAID 10 comparison.

The best for data backup is RAID 10, because it is more reliable due to URE and can withstand multiple disk failures, the rebuild is faster, the I/O performance is better and read performance is not degraded with disk failure.

Table 1.9: Comparison of RAID applicable for data backup.

	Advantages	**Disadvantages**
RAID 6	• More capacity	• Write performance penalty • Slower rebuild • Slower read (degraded)
RAID 10	• Better write performance • Faster rebuild • Can withstand failure of multiple disks (if it is not in the same mirror)	• Less capacity

If we want to save money and buy less disk to achieve higher capacity with good reliability, we can choose RAID 6. If we have only two or three drives, we can use RAID 1 to achieve better reliability and read performance.

Erasure coding

Due to the continual data growth, we have reached the limit of RAID solutions both due to URE (mentioned in the previous subsection 1.4) and due to lengthy RAID rebuild in a large-scale data storage (hundreds of terabytes and more). For this reason, we started to use erasure coding [58].

Erasure coding is a parity based protection technique where the data is first split into n chunks, then coded into m parity blocks and finally stored in various locations and storage media [59]. The advantage is that it consumes less storage space than replication, conversely, the calculation of parity is CPU-intensive and it increases latency.

The technique is widely used by cloud providers, who are looking to enhance data protection and durability. It can be used by medium and large companies with huge amounts of data for applications or systems with the need to tolerate failures, e.g. object storage, big data or archival storage.

Comparison of data storage medium

In the following table 1.10, general backup storage media in terms of life span, advantages and disadvantages is compared.

In the following table 1.11, concrete media representative with key factors such as: capacity, data rate and price per GB are presented. Storage is listed in order from hot storage (demands high performance) to warm and cold storage which can maintain large quantities of data at a relatively low cost with long-term reliability. The values were taken from current products on

Table 1.10: Comparison of general storage media.

Storage medium	**Life span**[1]	**Advantages**	**Disadvantages**
SSD	5 – 10 years	• The best access time • Less power draw than HDD • No moving parts	• The highest cost/GB • High rate of uncorrectable errors [57]
HDD	3 – 5 years[2]	• Low access time • Low cost/GB	• High power consumption • Producing heat
Tape	30 years	• Lowest cost/GB • Power-efficient • Stability	• High access time (load time and random access) • Require additional tape driver • Compatibility of LTO driver
Optical	100 years	• Longevity • Power-efficient • Durability • Stability • Backward compatible	• Moderate access time (load and latency of the head) • Moderate cost/GB • Require special disc drive • Traditional CD, DVD and Blu-ray disks are slow

[1] The life span expectancy for the tape media is up to 30 years in a well-maintained storage environment, but in real environment the life span could be far lower around 5-10 years, it depends on the frequency of using tape and environmental conditions.
[2] The reliability of HDD after 3-5 years of usage is rapidly declining, so we should consider a replacement during that time [60], [61].

the market in March 2017. If there were more products of same quality, then the average value was taken.

Summary

SSD is a storage medium used mainly for primary storage, with the trend of constantly decreasing price and the requirement to decrease RTO, it becomes in demand for backup usage as well. Besides backup storage HDD begins to be used as an archive storage replacing the original tape and optical archive storage media.

Some argue, that the tape is a dead technology and using cloud or HDD is better, but if we look to the table 1.11, we can see, that the tape has the best price per GB with a comparable performance. Although, there are more

Table 1.11: Comparison of specific storage media.

Storage class	**Model**	**Capacity**	**Write performance**	**Price per GB**
SSD	Intel DC S3520	1.6 TB[1]	380 MB/s	$0.475
HDD	Ultrastar He12	12 TB	255 MB/s	$0.031
	Seagate Archive	8 TB	190 MB/s	$0.03
Tape	LTO 6	2.5 TB	160 MB/s	$0.01
	LTO 7	6 TB	300 MB/s	$0.018
	TS1150	10 TB	360 MB/s	N/A
Optical	ODC3300R	3.3 TB	250 MB/s	$0.05
	M-disk	100 GB	18 MB/s	$0.2

[1] SSD is currently storage medium with largest capacity available on the market.

factors to consider, we should still count on the tape as Google did in 2011 when it was used for the recovery of Gmail data [62].

The idea that the longevity of optical media is ideal for archiving, tried to restart Facebook [63]. Sony continued with the idea and invented Everspan. John Fruehe from Moor Insights & Strategy in the paper [64] described the suitability of this solution for archiving by the comparison of the optical archive with tape archive in term of longevity and HDD archive in terms of cost and capacity.

As we can see, each of these storage mediums has its advantage and disadvantages so if we want to get the best from the backup solution and our budget allows it, we should use a combination of them all.

Tiered data backup storage strategies

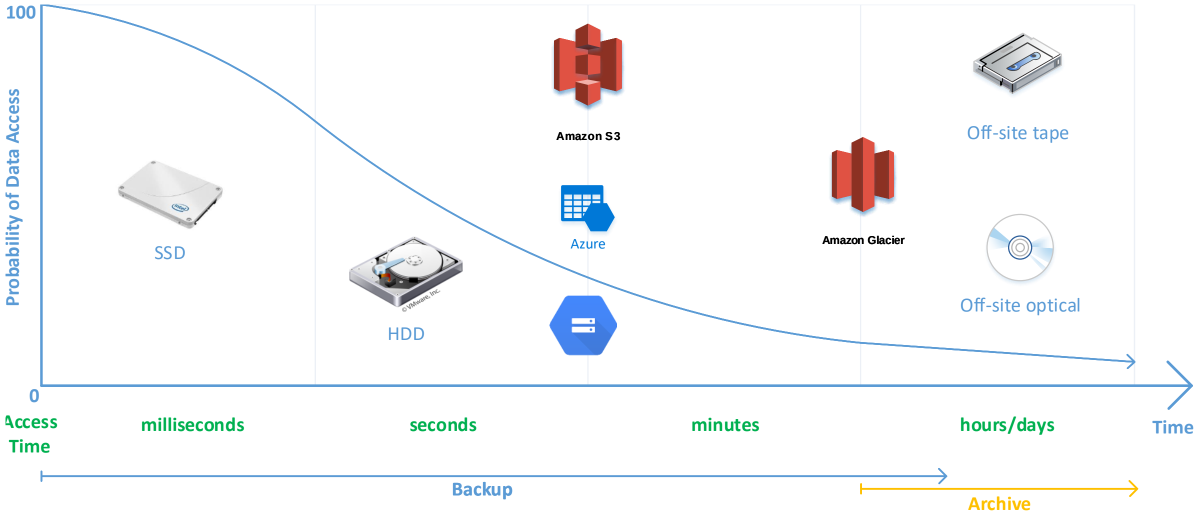
The ideal storage for the backup meeting the following attributes:

- Infinite capacity
- Zero cost both in term of purchase cost and cost of maintenance.
- Unlimited speed
- Endless Read/Write ability
- Completely reliable
- Infinity life span

Does not exist. Therefore, to get closer to the ideal storage we must combine the advantages of available storage on the market. That is called tiered storage. The amount of backup tiers depends on many factors, e.g. our business requirements and goals, our budget, manpower etc.

In the following picture 1.2, we can see how dividing the tiers might look like. The fresh backup data is where the highest probability of its use occurs, so it is stored on the fast medium with the short access time (SSD and HDD). Over the time, the data are moved to a slower and cheaper medium (HDD, Cloud storage and tape). In the end, the backup data are moved to the cheapest archival media (Glacier, tape and optical).

Figure 1.2: Tiered backup storage from the perspective of access time.



1.5 Calculation of storage system availability

In this subsection, the calculation of local storage availability will be introduced. This will serve for the comparison between our local storage and the cloud storage availability.

Availability of a single component

Theoretical availability of a single component can be calculated by the following equation 1.3 [65]. The reliability parameter Mean Time Between Failures (MTBF) can be found in the manufacturer's specification of the product. Mean Time To Repair (MTTR) can be calculated from historical data ($\frac{totaldowntime}{numberofrepairs}$), or we can estimate its value (e.g. We have a cold spare, first we have to find out that our storage is down, then we or our colleges have to replace the failed disk which may take on an average of 6 hours. Second case, we have a service agreement with HW provider, so we can use the number of hours in the agreement.).

$$Availibility = \frac{MTBF}{(MTBF + MTTR)} \tag{1.3}$$

However, the MTBF value does not often correspond to the real environment [60] and [66]. For a more accurate calculation, we can use another reliability parameter, namely the Annualized Failure Rate (AFR). The advantage of this value is that it can be found on the internet from the real environment [67], so it is no longer just theoretical. AFR can be approximately transferred to MTBF by the following equation 1.4. This equation assumes that the drives are powered on for the full 8760 hours of a year.

$$MTBF = \frac{8760}{AFR} \tag{1.4}$$

The availability increase of our system can be achieved either by increasing MTBF (more reliable disks, RAID configuration or system redundancy) or decreasing MTTR (better service conditions, hot or cold spare).

Availability of a multiple components

From the previous subsection, we know how to calculate the availability of a single component (disk). In this subsection, the calculation of more disks connected into one storage system will be introduced. There are two basic ways to link the system components together, serial and parallel configuration.

In a serial configuration, if any of the component in the series fails then the whole system fails (representative is RAID 0). Availability of a system in serial configuration is shown in the equation 1.5 where the availability of an individual disk is used and multiplied to one another.

$$Availability(Serial) = \prod_{i=1}^{n} Avalability(disk_i) \tag{1.5}$$

Where:

- i: represents the disk number
- n: represents the total number of disks [68]

In a parallel configuration, with each new component the redundancy is added and the system is operational if at least one component is functional (representative is RAID 1). Availability of a system in parallel configuration is calculated like *1-all components are unavailable* (see equation 1.6).

$$Availability(Parallel) = 1 - \prod_{i=1}^{n}(1 - Avalability(disk_i)) \tag{1.6}$$

Where:

- i: represents the disk number
- n: represents the total number of disks

Or we can have a combination of both serial and parallel configuration. Typical representative is RAID 10 and its availability is calculated by the following equation 1.7.

$$Availability(RAID\ 10) = \prod_{i=1}^{n/m} (1 - (1 - \prod_{j=0}^{m-1} Avalability(disk_{i*2-1}))) \tag{1.7}$$

Where:

- i, j: represents the disk number
- n: represents the total number of disks
- m: number of disks in the mirror, typically two

Availability of a RAID with parity

Because of the problem with the rebuild of array with parity mentioned in section 1.4. We should use RAID 6 and expect that it can handle only one disk failure. The availability of such configuration can be calculated by the following equation 1.8.

$$\begin{gathered} Availability(RAID\ 6) = P(X \leq 1) = P(X = 0) + P(X = 1) = \\ \prod_{i=1}^{n} Avalability(disk_i) + \sum_{i=1}^{n}((1 - Avalability(disk_i)) * \prod_{j=1;j \neq i}^{n} Avalability(disk_j)) \end{gathered} \tag{1.8}$$

Where:

- i, j: represents the disk number
- n: represents the total number of disks
- X: number of failed disks

1.6 Calculation of storage system durability

In this subsection, the calculation of storage durability will be introduced. This will serve as a comparison between our local storage and the cloud storage durability. The following formulas 1.9 and 1.10 were taken from the public durability calculator for Swift object storage [69], but can also be used to calculate the durability of the local storage with minor changes. The durability calculation of the local storage with parity RAID is shown in formula 1.9 and the durability of RAID 1 or RAID 10 is shown in formula 1.10.

$$\begin{aligned}Durability =& 1-(d*AFR*(1-AFR)^{d-1}+(1-AFR)^{c}*URE*8*s)*\\ & \prod_{i=1}^{c-1}((d-i)*RFR*(1-RFR)^{d-1-i}+(1-RFR)^{c-i}*URE*8*s)\end{aligned} \tag{1.9}$$

$$\begin{aligned}Durability =& 1-(d*AFR*(1-AFR)^{d-1}+(1-AFR)^{c}*URE*8*s)*\\ & \prod_{i=1}^{c-1}(i*RFR*(1-RFR)^{i-1}+(1-RFR)^{i}*URE*8*s)\end{aligned} \tag{1.10}$$

Where:

- c: Number of data Copies (number of mirrors, or parity disks + 1).
- d: Number of Disks.
- s: Size of the backup file in bytes.
- AFR: Annual Failure Rate. Probability of a disk failure in one year.
- RFR: Rebuild Failure Rate. Probability of a disk failure during rebuild. Calculated as $AFR*\frac{t_{rebuild}}{8760}$ ($t_{rebuild}$ - rebuild time in hours; 8760 - hours in a year).
- URE: Probability of an unrecoverable error per bit read.

1.7 Backup recommendation and strategy

The main reason to perform a data backup is because some time in the future we would need to recover the data we lost. Because of that, the most important part of our backup strategy is to test our backup regularly and check if we can recover our lost data and fulfill our planned goals as RTO.

In this chapter, we will see how to backup, what to backup and how often to do a data backup.

The 3-2-1 Rule

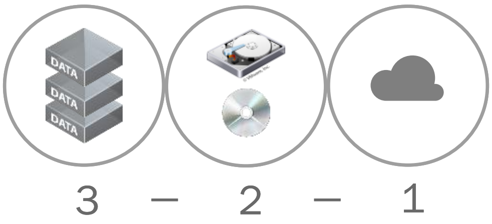
A good way to back up our data and increase our chances of recovery lost or corrupted data is to follow the rule 3-2-1 (see the following image 1.3)[70]. Where:

3 – Have at least 3 copies of any important file (a primary and two backups).

2 – Keep the files on 2 different media types, to protect against different types of hazards.

1 – Store 1 copy off-site, to protect against disaster.

Figure 1.3: Backup strategy 3-2-1.



This rule should be applied to everyone who has important data to backup and wants to protect them. It is the minimum configuration of proper backup, which could be expanded.

What to backup

We should backup everything, due to a quick and easy data recovery or restore. In case we do not have enough financial resources to back up everything, we can analyze our data and prioritize them from the most important to the least important data for our business. After prioritizing, the data backup will be concentrated first on the highest priority data containing critical data which is difficult to restore without proper backup (e.g. emails, production databases, financial and customer data) then on medium priority data (e.g. developer servers) and finally on the lowest priority data (e.g. personal data and publicly available data).

Backup schedule

How often to do data backup depends on business requirements such as RPO (cost of RPO should be lower than cost of data loss, for more information how to estimate cost of RPO and cost of data loss see 2.4). RPO is assigned to applications based on the importance/priority of the data and how often the data is changed. Backup is recommended to be scheduled for a time when our system resources and network usage is the lowest. Typically, this occurs overnight and in that case, our RPO is one day. If there is a requirement to do a backup during business hours, it is possible that the backup can affect our system or network resources even though today's backup tools are enabled to adapt the backup accordingly to the workload of our system or network resources.

1.8 Summary

In the previous chapter, a theoretical background of data backup and archiving was introduced. The information from the basic storage media, backup types and cloud storage to advanced techniques and strategies used in the data backup with the availability of backup storage were presented. Mentioned information will serve in the following chapter 2 as a starting point for cost-benefit analysis.

Chapter **2**

Cost-benefit analysis and design of hybrid cloud backup solutions

In this chapter, will be performed a cost-benefit analysis of hybrid cloud backup for SMEs. Consequently, will be designed a mathematical model, used to evaluate various hybrid backup solutions based on cost-benefit analysis. This is due to the creation of general comparison tool, in order to meet the objective. The objective is to create a set of general recommendations for the backup in SMEs applying both technological and managerial dimension terms.

First will be introduced the cost and benefit factors of backup solutions. Second will be presented a mathematical model. Third will be described how the individual components of the model are calculated. Fourth will be created a tool for calculation of mathematical model. Then will be utilized the model to different types of businesses within SMEs using different types of backup solutions and the tool will be used for evaluating the precise numbers. Finally, based on the previous results, will be created a set of general recommendations for the backup in SMEs applying both technological and managerial dimension terms.

2.1 The scope of the analysis

The analysis includes the following items:

- Small and medium-sized enterprises
- Public cloud backup storage within three CPS leaders:
  - Amazon Web Services
  - Google Cloud Storage
  - Azure Storage

- Data backup on a storage media such as SSD and HDD

The analysis excludes the followings items:

- Private cloud: because building our own private cloud, using managed hosting, or colocation environment is not suitable for SMEs that wants an off-site environment used for data backup [2], [71], [72] and [73].
- Optic media: because optical medium with WORM technology is designed mainly for archiving.
- Tape media: because of the higher price of tape library it is not suitable for small companies.

2.2 The cost and benefit factors of backup solutions

Cost factors

All cost factors mentioned in the analysis of the data center [1] and [2] can also be apply for SMEs. However, some of these factors are meaningless due to a small value. In contrast, new cloud cost factors were added because of the new pricing models by CSP (AWS, Google and Azure) which are constantly changing.

Prerequisites for the analysis assumes that the company already has a basic IT equipment, including production servers, the underlying network infrastructure (switches and routers) and cooling (air conditioner) in a dedicated area. Therefore, there is no need to buy new network devices due to one or a few additional devices on the network (backup storage), air conditioners or facility.

In that case, the cost factors could be categorized into five groups. Two of them, hardware and electricity are on-premises cost factors. The third cloud service is a cloud cost factor and the last two software and labor belongs to both on-premises and cloud cost factors (see the following table 2.1).

Hardware: is the most expensive group in on-premises. It consists purchase price of the storage server, disks and additional HW, such as uninterruptible power supply (UPS) or network components. Disks are separated due to the great acquisition cost and in some network-attached storage (NAS), we must purchase them separately.

Electricity: The power usage in the SMEs is not that significant as in data center cases, but it will also be included into the analysis to not favor a local solution compared to the cloud. Electricity consumption is divided into two groups: electrical consumption for electronic devices and for cooling those electronic devices which produce heat.

Labor: comprises wages for IT specialist who work mainly on maintaining hardware, but also software. The cost of maintenance SW falls under both on-premises and cloud category (e.g. operating system (OS) of the storage server is on-premises, WAN optimization is cloud while backup and restore SW falls under both categories).

Software: includes the purchase price of software and additional licenses, especially the backup and restore software, but the missing OS and WAN optimization SW can also be included. This group falls under both on-premises and cloud cost factors depending on the type of software.

Cloud service: Cost factors for cloud backup storage are not just about choosing a storage class with its capacity, but it includes other additional cost factors that are tied to the cloud storage and its use. Second cost factor emerges from network usage when we transfer the data out of the cloud storage (egress) into our own local storage or different CSP, in some cases there are even charges for data transfer within the same CSP. Other cost factors besides cloud storage are operations performed in cloud storage. Within the three leading CPS are those operations divided to three groups, one of them associated with delete request is free of charge. For generalization, the other two groups are named class operations A (with a representative PUT) and class operations B (with a representative GET). The operation falling within these groups vary depending on the CSP. Regarding cool and cold data storage, extra operation fees are added because of the lower price for storage and a commitment to keep the data in the storage for a specified amount of time. We incur additional cost factors for data access namely data retrieval and early deletion. Another cost factor for cool Azure storage only is data write. Furthermore, other price factors are the data transfer acceleration in and out of the CSP. AWS offers this service directly, or we can use any third-party WAN optimization software which is integrated with our CSP. Another way to transfer especially large amount of data on offline media is import/export service. AWS provides a service called AWS Snowball, Azure offers Import/Export, as for Google import/export we can use a third-party service provider. Finally, the need for more bandwidth can arise with the use of cloud services. So, the last cost factor is an additional cost for internet connectivity.

Benefit factors

Benefit factors of data backup storage were found from the analysis of parameters in each individual media type to data backup storage analysis as a whole from the point of view of both local and remote solutions. The benefit factors are: performance, availability, durability, reliability, security, scalability, capacity, off-site data protection, manageability, accessibility and provided features.

Performance: Data storage performance is composed of the access time to data/storage (latency), data transfer rate (throughput) and time needed by

Table 2.1: Overview of cost factors.

Cost factor scope	Group	Cost factors
On-premises	Hardware	Storage server Disks Additional HW
	Electricity	Electronic devices Cooling
On-premises and cloud	Labor	HW maintenance SW maintenance
	Software	Software
Cloud	Cloud service	Cloud storage Class A operations Class B operations Data transfer out (egress) Data retrieval Early deletion Data write Data transfer in acceleration Data transfer out acceleration Import Export Internet connectivity

storage to perform both input and output operations (IOPS).

Availability, durability and reliability: The higher reliability parameters is the higher availability and durability our system has. The difference between availability and durability is that availability is given by a number of redundant HW components and durability is given by a number of copies of our data [74].

Security: is divided into two groups. Physical security of the facility, or a room where the data is stored and the way the data is stored or transfer. The question of whether our data is encrypted with a strong secure cipher is especially important when storing data remotely over the unprotected public network. To check the security of service providers we can look at provider compliance to see which security standards the provider meets (e.g. ISO 27001, ISO 27017, ISO 27018).

Scalability: In the case of scalability of the data backup system there are two questions to consider. Is it possible to scale our data backup storage up, or out? How difficult is it to scale our data backup storage? Where is scale out more important, because if we do not use complex optimization techniques such as data deduplication, we do not need to scale up, but over time as the data to be backed up is growing the need to scale out is likely to arise.

Off-site data protection (vaulting): is the benefit of storing data off-site in a secure facility to protect our local data against disaster.

Manageability: determines how easy it is to manage or control our data backup storage. For comparison, we can ask the following questions. Do I need additional knowledge to access or control the data backup storage with a backup software or via console? How difficult is it to set up and maintain the data backup in the backup management? How efficiently can I monitor the data backup storage?

Accessibility: determines the accessibility of our data backup solution (e.g. We can access the storage from everywhere, home, or just from work network).

Provided features: is a list of features the data backup storage or backup software provides. It can be the integration of well-known service providers or our existing architecture, a level of automation (manual vs automatic backup), offered backup options (incremental, full, synthetic), backup optimization (data compression, data deduplication, transfer acceleration), graphical interface or encryption.

Evaluation of cost-benefit factors

In this subsection, individual cost and benefit factors will be evaluated from the point of view of SME. The evaluation of cost-benefit factors is important for the comparison between local, cloud and hybrid data backup solutions. For this, all parameters must be converted to a common unit. As the common unit (currency), the US dollar was chosen because it is the most used currency in the world and the thesis is written in English.

The evaluation of cost factors is trivial. Cost factors can be found in US dollars, so for comparison, we will only take the value, add it to the total cost and finally, compare the whole. The only problem is that certain factors are associated with a qualified estimate (e.g. storage usage in terms of operations).

The evaluation of benefit factors is more complex, considering that individual benefits are usually only parts of the purchased data backup solution, so we cannot take the total cost of it. Therefore, in this subsection, we will try to examine the individual benefit parameters and evaluate their importance in terms of involvement in the mathematical model in the following section 2.3.

The most important benefit parameters in the mathematical model includes: performance, availability and durability. Performance parameters are directly linked to the disaster recovery (DR) parameters RPO and RTO. These parameters are well known DR parameters which allow us to compare the individual backup solutions. The comparison of individual performance parameters is pointless because the overall performance is governed by the slowest part of the system (bottleneck). Availability is another important parameter.

Our backup storage should be available whenever we need to access it, otherwise, our RPO and RTO will increase. Finally, durability comes in line when we access the data expecting that no data will be damaged or lost. How to calculate these benefit parameters will be shown in the following section 2.4.

Other benefit parameters: security, scalability, accessibility and off-site data protection are important as well, so why not include these parameters in the mathematical model? First, these parameters are difficult to evaluate and second, there is no reason to do that. In the case of security, certified CSPs always provide a higher level of physical security than SMEs, but there is no need for a higher security level of data backup in SMEs when our production data does not reach the same level of security. Meanwhile, encryption of data to communicate with the remote facility is not an advantage but a standard. Scalability is again simpler and easier in cloud solutions, but limited scalability offered by local solutions are sufficient for SMEs needs. In accessibility, both solutions cloud and local can be accessed from anywhere if required. Off-site data protection is an important benefit factor that we should look for if we want to meet the backup recommendation (see the 3-2-1 rule in section 1.7). This is another benefit of cloud solutions, and in case we only have on-site backup solution we should consider simultaneously storing the data off-site.

The last two benefit factors: manageability and provided feature are more connected with SW than with the backup solution since their evaluation is reflected in cost factor maintenance, so we do not need to evaluate them separately.

The cloud solutions are superior in most of the benefit factors, however, in most cases we can achieve the desired goal even with local solutions. This suggests that if we do not find an unambiguous winner, we should prefer the cloud solution over the local solution.

2.3 The cost-benefit model design

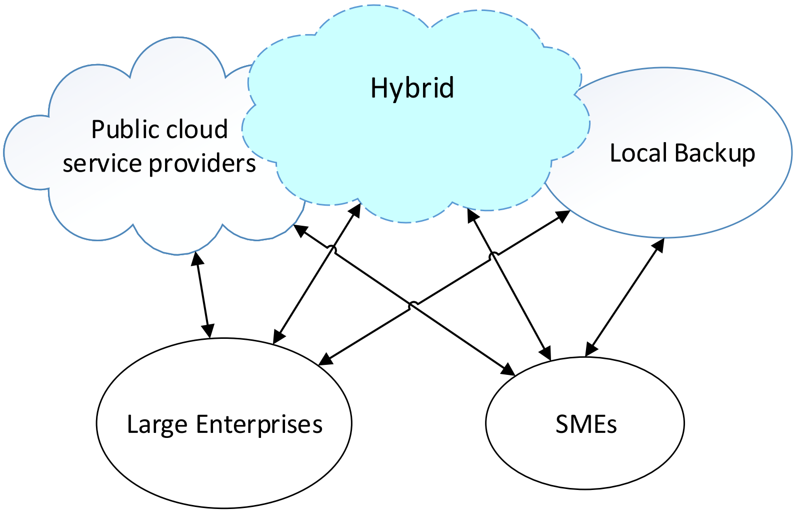
Model context

The context of our mathematical model considers all types of companies from SMEs to large enterprises that are thinking about the optimal backup solution which may be either a local, cloud or hybrid solution as shown in the following figure 2.1.

Mathematical model

The proposed mathematical model is a type of nonlinear optimization that is used to find the optimal backup solution. The optimal solution is a solution where cost is minimal and parameters are meeting the requirements (goals) specified by the SMEs. This relationship was chosen because the price is often the most important decision making factor, yet also includes qualitative factors. The nonlinearity is given by benefit parameters that prevails for either local or cloud solutions.

Figure 2.1: The context of mathematical model.



The objective function of the nonlinear program is worded as follows:

Minimize Cost for certain Availability, RPO, RTO and Durability.

Mathematical notation is shown in the following formula 2.1. And graphical interpretation with the indicated feasible region is shown in the following figure 2.2.

Objective function:

$$\begin{aligned}
&Minimize: \sum_{i=1}^{m} CL_i x_i + \sum_{j=1}^{n} CC_j y_j + \sum_{i=1}^{m}\sum_{j=1}^{n} (L_{ij} z_{ij})\\
&Subject\ to:\\
&\qquad \sum_{i=1}^{m} CL_i x_i + \sum_{j=1}^{n} CC_j y_j + \sum_{i=1}^{m}\sum_{j=1}^{n} (L_{ij} z_{ij}) \leq B\\
&\qquad 0 \leq C_{RPO} \leq (C_{DL} \wedge B)\\
&\qquad 0 \leq C_{RTO} \leq (C_{SD} \wedge B)\\
&\qquad 0 \leq C_A \leq B\\
&\qquad 0 \leq C_D \leq B\\
&\qquad CL_i \geq 0,\ CC_j \geq 0,\ L_{ij} \geq 0 \qquad ; 1 \leq i \leq m, 1 \leq j \leq n\\
&\qquad x_i, y_j, z_{ij} \in \{0,1\}
\end{aligned} \tag{2.1}$$

Where:

- CL_i: Cost of Local backup system (see next section 2.4).

- CC_j: Cost of Cloud backup system (see next section 2.4).
- C_{DL}: Cost of Data Loss (see next section 2.4).
- C_{SD}: Cost of System Downtime (see next section 2.4).
- C_{RPO}: Cost of RPO (see next section 2.4).
- C_{RTO}: Cost of RTO (see next section 2.4).
- C_A: Cost of Availability (see next section 2.4).
- C_D: Cost of Durability (see next section 2.4).
- L_{ij}: The link cost between (i, j), i, j $\in N$.
- x_i: The binary variable indicating whether at i is a local backup solution or not.
- y_j: The binary variable indicating whether at j is a cloud backup solution or not.
- z_{ij}: The binary variable indicating whether is dedicated link between our building at i and cloud at j or not.
- B: Our Budget, dedicated money for backup solution.

2.4 Description of cost-benefit model

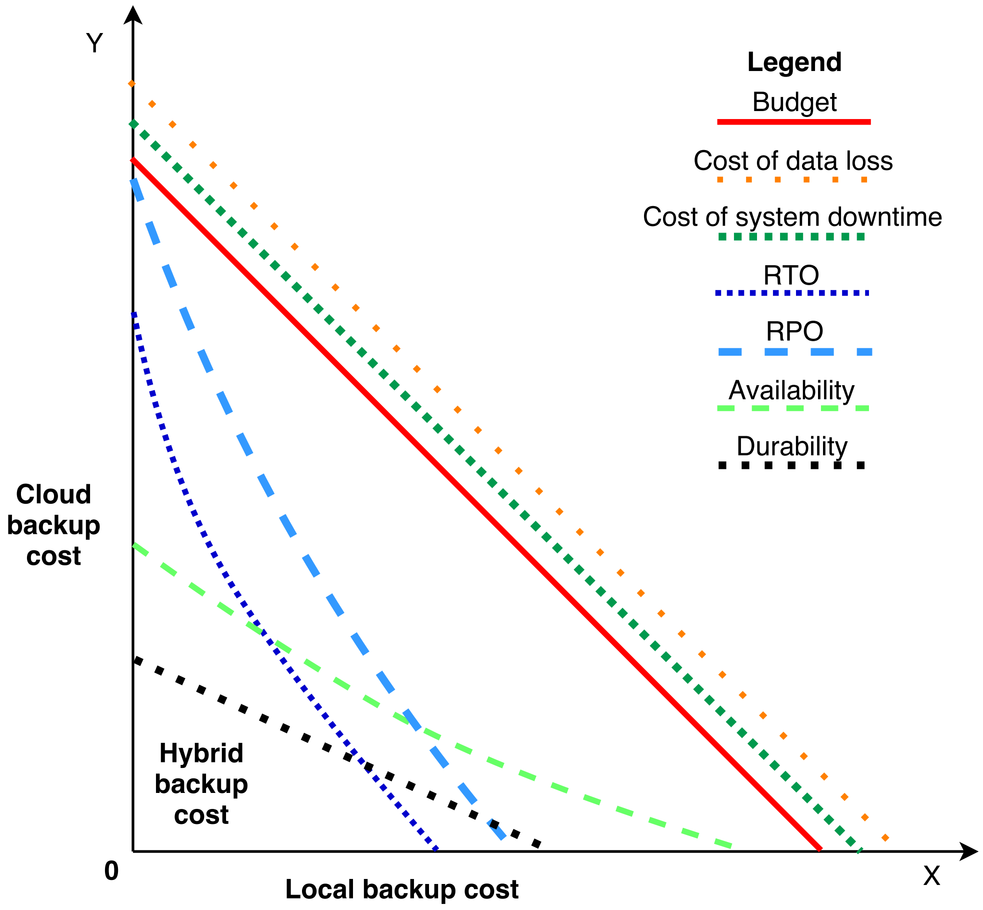
In this subsection, we will see methods on how to simply approximate the calculation of each individual component from the model 2.3. These methods are intended for general use. The methods are approximate because there are too many variables in total within the variety of companies.

Cost of Local and Cloud backup system

The cost of the local backup system is calculated by the sum of all the individual items belonging to cost factors within on-premises scope (see table 2.1).

The cost of the cloud backup system is calculated by the sum of all the individual items belonging to the cost factors within cloud scope (see table 2.1).

Figure 2.2: The graphical interpretation of mathematical model.



Cost of system downtime

According to the research report [75], the cost of system downtime consists of nine categories, ordered from the most expensive category (Business disruption) to the least expensive category (Third parties):

- Business disruption: The cost associated with business disruption, which includes reputation damages and customer churn, represents the most expensive cost category.

- Lost revenue: The total revenue loss from customers and potential customers because of their inability to access core systems during the outage period.

- End-user productivity: The lost time and related expenses associated with end-user downtime.
- IT productivity: The lost time and related expenses associated with IT personnel downtime.
- Detection: Activities associated with the initial discovery and subsequent investigation of the partial or complete outage incident.
- Recovery: Activities and associated costs that relate to bringing the organization's networks and core system back to a state of readiness.
- Equipment repair & replacement: The cost of equipment new purchases and repairs, including refurbishment.
- Ex-post activities: All after-the-fact incidental costs associated with business disruption and recovery.
- Third parties: The cost of contractors, consultants, auditors and other specialists engaged to help resolve unplanned outage.

Calculation the cost of each category

It is not easy to find out the true cost of **business disruption** (consequences), for that we would need a long history of data and still it would be inaccurate. But since it is the most valuable category impacting the overall cost of system downtime in certain cases, we should make at least a rough estimate. Remember that if our company is oriented in an industry segment, such as communication, e-commerce or banking, then this company will be heavily affected by this category.

Estimation of the total cost in this category consists of **lost customers** and **lost reputation** (see equation 2.2).

$$\begin{aligned} Cost\ of\ lost\ customers &= T_{NC} * L_{VC} * I_{LC} \\ Cost\ of\ lost\ reputation &= T_{PC} * L_{VC} * I_{LR} \\ Cost\ of\ business\ disruption &= L_{VC} * (T_{NC} * I_{LC} + T_{PC} * I_{LR}) \end{aligned} \tag{2.2}$$

Where:

- T_{NC}: Total Number of Customers.
- T_{PC}: Total number of Potential Customers.
- L_{VC}: Lifetime Value of Customer (e.g. how much money a customer spends in our e-shop).
- I_{LC}: Impact on the Loss of Customers in percentage (Percentage of customers who are permanently lost, e.g. their loyalty is low, so they go to the competitor.)

- I_{LR}: Impact on the Loss of Reputation in percentage.

As we can see, these values are mostly just educated guesses or plausible ranges, except the total number of customers and their lifetime value, which we can get from our long history of data.

Lost revenue depends a lot on how your organization makes money and is estimated by the following equation 2.3.

$$\begin{aligned} Revenue\ cost\ per\ day &= average\ revenue\ per\ day * I_R \\ &= (\frac{GR}{T_{BD}}) * I_R \\ T_{BD} &= business\ days\ per\ week * weeks\ per\ year \end{aligned} \tag{2.3}$$

Where:

- GR: Gross yearly Revenue.
- T_{BD}: Total yearly Business Days.
- L_{VC}: Lifetime Value of Customer (e.g. how much money a customer spends in our e-shop).
- I_R: Revenue Impact in percentage (Percentage of Revenue affected by downtime). Its value is just educated guess or plausible range.

Revenue impact is affected by business orientation. If our business is project-oriented, we are not that dependent on server uptime (the value is closer to 0 %). On the other hand, e-commerce business is highly dependent on server uptime (the value is closer to 100 %).

If the case is e-commerce usually running 24/7, then the T_{BD} is regularly 365 (days). Consider that most orders may arrive in a certain time range (8-14 hours), depending on the business region (international or national).

End-user productivity and **IT productivity** will be included in one group called **Lost productivity** (see the equation 2.4), because nowadays most people work with computers and information systems so both categories have common characteristics.

$$\begin{aligned} Productivity\ cost\ per\ day &= T_{NE} * employee\ cost\ per\ day * I_E \\ &= (\frac{total\ cost\ of\ employee\ salaries\ and\ benefits\ per\ year}{T_{WD}}) * I_E \\ &= T_{NE} * (\frac{E_{SB}}{T_{WD}}) * I_E \\ T_{WD} &= working\ days\ per\ week * weeks\ per\ year \end{aligned} \tag{2.4}$$

Where:

- E_{SB}: Average Employee Salary with Benefits per year.

- T_{NE}: Total Number of Employees dependent on IT system.
- T_{WD}: Total yearly Working Days.
- I_E: Employee Impact in percentage (Percentage of Employees affected by downtime). Its value is just educated guess or plausible range.

The remaining categories (activities) such as **detection, recovery, equipment repair & replacement, ex-post activities** and **third parties** will be included in one group called **Cost to recover** (see the equation 2.5). This is due to reason that in SMEs most of these activities are done by one man or one single IT department or resolved by a third party (outsourced).

$$Cost\ to\ recover = N_{IE} * E_{IDBR} * T_{LI} + N_{EE} * E_{EDBR} * T_{LE} \tag{2.5}$$

Where:

- E_{IDBR}: Average cost of Internal Employee who is responsible for Data Backup and Recovery per hour.
- E_{EDBR}: Average cost of External Employee who is responsible for Data Backup and Recovery per hour.
- N_{IE}: Number of Internal Employees responsible for system recovery.
- N_{EE}: Number of External Employees responsible for system recovery.
- T_{LI}: Total Labor hours of an Internal employee engaged in IT required to detect and recover the system.
- T_{LE}: Total Labor hours of an External employee engaged in IT required to detect and recover the system.

Cost of the system downtime If we put together the previous equations 2.2, 2.3, 2.4 and 2.5, we will get the following estimate for the cost of the system downtime, which tends to be conservative. The first part of the equation 2.6 is in days, since we do not calculate non-working/business hours facing system downtime.

$$\begin{aligned} C_{SD} =& (revenue\ cost\ per\ day + productivity\ cost\ per\ day) * \\ & D_{SD} + cost\ to\ recover + cost\ of\ business\ disruption \\ =& ((\frac{GR}{T_{BD}}) * I_R + T_{NE} * \frac{E_{SB}}{T_{WD}} * I_E) * D_{SD} + N_{IE} * E_{IDBR} * T_{LI} + \\ & N_{EE} * E_{EDBR} * T_{LE} + L_{VC} * (T_{NC} * I_{LC} + T_{PC} * I_{LR}) \end{aligned} \tag{2.6}$$

Where:

- C_{SD}: Cost of the System Downtime.
- D_{SD}: Duration of System Downtime in days (to use different units, just divide the number by the exact constant to get a day, $\frac{hours}{24}$; $\frac{minutes}{1440}$; $\frac{seconds}{86400}$).
- E_{IDBR}: Average cost of Internal Employee who is responsible for Data Backup and Recovery per hour.
- E_{EDBR}: Average cost of External Employee who is responsible for Data Backup and Recovery per hour.
- E_{SB}: Average Employee Salary with Benefits per year.
- GR: Gross yearly Revenue.
- I_E: Employee Impact in percentage (Percentage of Employees affected by downtime). Its value is just educated guess or plausible range.
- I_{LC}: Impact on the Loss of Customers in percentage (Percentage of customers who are permanently lost, e.g. their loyalty is low, so they go to the competitor.)
- I_{LR}: Impact on the Loss of Reputation in percentage.
- I_R: Revenue Impact in percentage (Percentage of Revenue affected by downtime). Its value is just educated guess or plausible range.
- L_{VC}: Lifetime Value of Customer (e.g. how much money a customer spends in our e-shop).
- N_{IE}: Number of Internal Employees responsible for system recovery.
- N_{EE}: Number of External Employees responsible for system recovery.
- T_{BD}: Total yearly Business Days.
- T_{LI}: Total Labor hours of an Internal employee engaged in IT required to detect and recover the system.
- T_{LE}: Total Labor hours of an External employee engaged in IT required to detect and recover the system.
- T_{NE}: Total Number of Employees dependent on IT system.
- T_{NC}: Total Number of Customers.
- T_{PC}: Total number of Potential Customers.
- T_{WD}: Total yearly Working Days.

Cost of data loss

Assume that everyone makes at least some kind of data backup, because if there were any unexpected disasters with no data backup, it would be the end of our company.

The cost of data loss has the same part of an equation as the cost of system downtime (see previous subsection 2.4) with a minor modification. The equation consists of **lost revenue, lost productivity, a cost to recover** and **data loss**.

The loss of revenue could arise due to the missing part of the database with goods, so in our e-shop, there will be some missing items that customers cannot buy thus our revenue is lost.

The loss of productivity in the workplace is due to waiting for data to be recovered from the data backup.

If we are using a public cloud for data backup, we must extend the cost to recover (in the case of data loss, the cost to data restore) with the payment for data transferred from the cloud (outbound traffic).

Data loss

Data loss is a new part of the equation. Data are divided into two groups: non-restorable and restorable. The restoration cost is mentioned in the previous subsection 2.4. In this section, the estimation of price non-restorable data will be introduced.

According to the report [76], data in our company can be divided into four categories: corporate data, customer data, transaction data and personally identifiable data. I will divide those data into two categories: data we can rebuild (corporate data) and data we cannot rebuild or the rebuild is complicated (customer data, transaction data and personally identifiable data).

Data we can rebuild are corporate data. Data created by employees, so if we lose those data we can recreate them again. The following equation 2.7 tells us how to estimate the cost of those data.

$$Cost\ of\ data\ rebuild = T_{NE} * (\frac{E_{SB}}{T_{WD}}) * I_E * T_{LB} \tag{2.7}$$

Where:

- E_{SB}: Average Employee Salary with Benefits per year.
- T_{NE}: Total Number of Employees dependent on IT system.
- T_{WD}: Total yearly Working Days.
- I_E: Employee Impact in percentage (Percentage of Employees affected by data loss). Its value is just educated guess or plausible range.

- T_{LB}: Time since the last backup in days (to use different units, just divide the number by the exact constant to get a day, $\frac{hours}{24}$; $\frac{minutes}{1440}$; $\frac{seconds}{86400}$).

Data we cannot rebuild are customer data, transaction data and personally identifiable data. Estimation of the total cost of those categories consists of data lost, lost customers and lost reputation. According to the report [77], the average global cost of a data breach per lost or stolen record was \$158 and the more records are lost, the higher the cost of the data breach is. Estimation of lost customers and lost reputation was presented in the previous subsection 2.4, the cost of data loss is shown in the following equations 2.8.

$$\begin{aligned} Cost\ of\ lost\ records &= C_{RL} * T_{RL} \\ Cost\ of\ lost\ customers &= T_{NC} * L_{VC} * I_{LC} \\ Cost\ of\ lost\ reputation &= T_{PC} * L_{VC} * I_{LR} \end{aligned} \tag{2.8}$$

Where:

- C_{RL}: Average cost of Record Lost.
- T_{RL}: Total Number of Record Lost.

Overall cost of data loss

By the connection of previous equations 2.7 and 2.8, we get the following estimation for the overall cost of data loss (see equation 2.9).

$$\begin{aligned} C_{DL} =&(revenue\ cost\ per\ day + productivity\ cost\ per\ day)* \\ &D_{DR} + cost\ to\ recover + cost\ of\ data\ lost \\ =&((\frac{GR}{T_{BD}}) * I_R + T_{NE} * \frac{E_{SB}}{T_{WD}} * I_E) * D_{DR} + N_{IE} * E_{IDBR} * T_{LI} + \\ &N_{EE} * E_{EDBR} * T_{LE} + A_{DT} * price\ per\ GB + \\ &T_{NE} * (\frac{E_{SB}}{T_{WD}}) * I_E * T_{LB} + C_{RL} * T_{RL} + L_{VC} * (T_{NC} * I_{LC} + T_{PC} * I_{LR}) \end{aligned} \tag{2.9}$$

Where:

- A_{DT}: Amount of Data Transferred from cloud to on-premises, other cloud service provider, or other zone in GB.
- C_{DL}: Cost of the Data Loss.
- C_{RL}: Average cost of Record Lost.
- D_{DR}: Duration of Data Recovery in days (to use different units, just divide the number by the exact constant to get a day, $\frac{hours}{24}$; $\frac{minutes}{1440}$; $\frac{seconds}{86400}$).

- E_{IDBR}: Average cost of Internal Employee who is responsible for Data Backup and Recovery per hour.
- E_{EDBR}: Average cost of External Employee who is responsible for Data Backup and Recovery per hour.
- E_{SB}: Average Employee Salary with Benefits per year.
- GR: Gross yearly Revenue.
- I_E: Employee Impact in percentage (Percentage of Employees affected by data loss). Its value is just educated guess or plausible range.
- I_{LC}: Impact on the Loss of Customers in percentage (Percentage of customers who are permanently lost, e.g. their loyalty is low, so they go to the competitor.)
- I_{LR}: Impact on the Loss of Reputation in percentage.
- I_R: Revenue Impact in percentage (Percentage of Revenue affected by data loss). Its value is just educated guess or plausible range.
- L_{VC}: Lifetime Value of Customer (e.g. how much money a customer spends in our e-shop).
- N_{IE}: Number of Internal Employees responsible for data backup and recovery.
- N_{EE}: Number of External Employees responsible for Data Backup and Recovery.
- T_{BD}: Total yearly Business Days.
- T_{LB}: Time since the last backup in days (to use different units, just divide the number by the exact constant to get a day, $\frac{hours}{24}$; $\frac{minutes}{1440}$; $\frac{seconds}{86400}$).
- T_{LI}: Total Labor hours of an Internal employee engaged in IT required to detect and restore the data.
- T_{LE}: Total Labor hours of an External employee engaged in IT required to detect and restore the data.
- T_{NE}: Total Number of Employees dependent on IT system.
- T_{NC}: Total Number of Customers.
- T_{PC}: Total number of Potential Customers.
- T_{RL}: Total Number of Record Lost.
- T_{WD}: Total yearly Working Days.

Cost of RTO

The cost of RTO is the cost required to achieve a certain RTO assuming that we do not have to buy a new equipment (component such as CPU, ram and disk of failed system). In that case, the cost of RTO is calculated by the following equation 2.10.

$$\begin{aligned} C_{RTO} =&C_{SC} + C_{BSW} + C_{EDR} + C_{DT} + C_{SR} + C_{ANR} + \\ &N_{IE} * E_{IDBR} * T_{LI} + N_{EE} * E_{EDBR} * T_{LE} \end{aligned} \tag{2.10}$$

Where:

- C_{ANR}: Cost of Additional Network Resources is important to achieve certain RTO (e.g. network bandwidth and WAN acceleration).
- C_{BSW}: Cost of Backup Software.
- C_{DT}: Cost of Data Transfer, charges associated with retrieving off-site data (e.g. in cloud it is outbound traffic = amount of data transferred from the cloud*price per GB, or shipping the storage media stored off-site).
- C_{SC}: Cost of data backup Storage with certain Capacity. It includes the cost of both local and cloud solution.
- C_{SR}: Cost of Storage Requests. In the case of cloud, we must pay for retrieval operations (GET operation).
- C_{RTO}: Cost of RTO.
- E_{IDBR}: Average cost of Internal Employee who is responsible for Data Backup and Recovery per hour.
- E_{EDBR}: Average cost of External Employee who is responsible for Data Backup and Recovery per hour.
- N_{IE}: Number of Internal Employees responsible for data backup and recovery.
- N_{EE}: Number of External Employees responsible for data backup and recovery.
- T_{LI}: Total Labor hours of an Internal employee engaged in IT required to detect and restore the data.
- T_{LE}: Total Labor hours of an External employee engaged in IT required to detect and restore the data.

Cost of RPO

The cost of RPO is defined as the cost required to achieve certain RPO. RPO is affected mainly by frequency of data backups, but also by bandwidth, backup storage performance and a load of our system. Decreasing traditional RPO with the value of one day, would lead to a backup during the working hours. For that we would need additional resources to not affect our employees who are working on a server that is backed up. Evaluating those additional resources can be challenging, but it is variable, which appears in both cases, local and cloud backup, so for the final comparison we can ignore this value. Lower RPO leads to a higher number of backups and hence to greater storage capacity needs. For off-site backup, it is possible that our internet connection is insufficient to transfer all data in a certain time to meet the RPO. In that case, we must invest to additional network resources and not forget the additional cloud fees for operations. When we put all these parameters together along with the price for the backup SW, we get the following equation 2.11.

$$C_{RPO} = C_{SC} + C_{BSW} + C_{SR} + C_{ASR} + C_{ANR} \tag{2.11}$$

Where:

- C_{ASR}: Cost of Additional System Resources for a non-backup environment, data backup storage is excluded (e.g. CPU and RAM).
- C_{ANR}: Cost of Additional Network Resources is important to achieve certain RPO (e.g. network bandwidth and WAN acceleration).
- C_{BSW}: Cost of Backup Software.
- C_{SC}: Cost of data backup Storage with certain Capacity. It includes the cost of both local and cloud solution.
- C_{SR}: Cost of Storage Requests. In the case of cloud, we must pay for an insertion of each object (PUT operation).
- C_{RPO}: Cost of RPO.

Cost of availability

The cost of availability will be defined as the cost required to achieve a certain availability (for more information on how to calculate system availability see section 1.5). This is calculated by the following equation 2.12.

$$C_A = C_{SC}(A) \tag{2.12}$$

Where:

- C_A: Cost of Availability.

- A: is the percentage of Availability.

- $C_{SC}(A)$: Cost of data backup Storage with certain Capacity achieving certain Availability. It includes the cost of both local and cloud solution.

Cost of durability

The cost of durability will be defined as the cost required to achieve a certain durability (for more information on how to calculate system durability see section 1.6).

Durability is affected primarily by the number of copies, the rebuild time and the reliability of the storage medium (AFR, URE). All of these parameters are related to the data storage and since the total number of copies affects the capacity of the media or the number of media, the cost of durability equals the cost of data backup storage with a certain capacity (see following equation 2.13).

$$C_D = C_{SC}(D) \tag{2.13}$$

Where:

- C_D: Cost of Durability.

- D: is the percentage of Durability.

- C_{SC}: Cost of data backup Storage with certain Capacity achieving certain Durability. It includes the cost of both local and cloud solution.

2.5 Data backup comparison tool

Excel tool named *Data_Backup_Comparison_Tool.xlsx* was created for mathematical model evaluation. This tool is designed to perform a cost-benefit analysis of data backup solutions for small and medium-sized enterprises. The tool is used to compare on-premises data backup solutions with cloud data backup solutions in terms of total cost of ownership and benefits arising from business impact analysis. The results from the analysis of on-premises and cloud data backup storage are combined into a hybrid data backup solution and can be used to compare the hybrid solution separately.

Each sheet in Excel serves to evaluate an individual parameter from the mathematical model (see formula 2.1). These individual results are then combined into the final sheet which shows an overview of all parameters along with graphical comparison and resulting recommendations.

2.6 Analysis of SMEs

For simplicity, I assume that business data is stored either on a shared storage or on an information system that runs on virtual servers. Therefore, it is important that the backup products support backing up shared storage and virtual servers. In the analysis, these data sources are merged into one and thus an important parameter: *the total size of the backed up data.* This parameter is important because the size of the company is not correlated to the total size the data company has. Therefore, it will be used to divide businesses within the SMEs.

All prices in Czech crowns during the calculation were converted to US dollars at the rate of 1 USD = 24 CZK.

Model Company 1

It is a medium-sized non-IT company with typical IT background (application servers, workstations, shared storages and databases) with the following information:

- Number of employees: 225
- Total size of the backed up data: 18 TB
- Internet connection speed: 1 000 Mb/s
- Data retention: 4 weeks
- RPO: 20 hours
- RTO: 36 hours
- Availability: 99.5%
- Durability: 99.99%

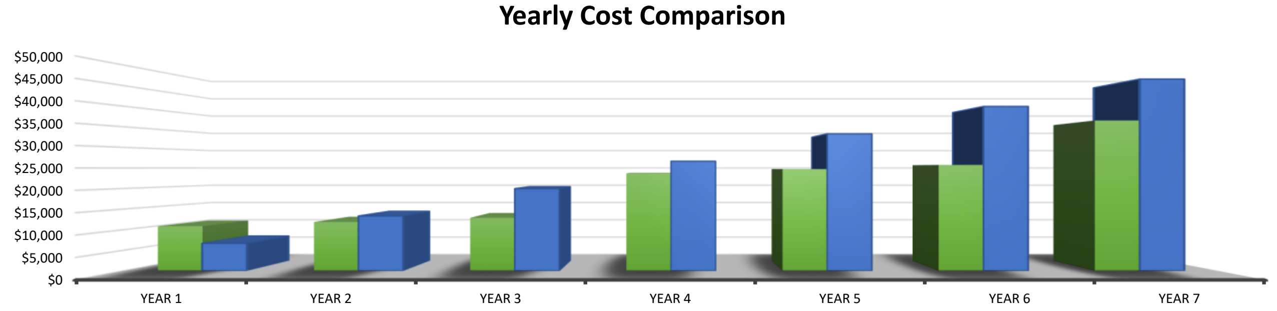
The total storage size was estimated to 66 TB, including compressed data, with no data deduplication, four incremental backups and one full backup a week with a ten percent data growth a year and three years of the refresh cycle.

For an on-premises storage server, we selected NAS DiskStation DS2415+, with a 6 TB Enterprise Capacity 3.5 HDD from Seagate in RAID 10 configuration, thanks to its many disks and large storage size. As a cloud storage class we selected Nearline Storage.

Although data growth and retrieval cost, which is the most expensive item in cloud storage, was not included in the cloud solution's overall price the on-premises data backup solution still clearly won the price in comparison with the cloud (see figure 2.3). The only way cloud backup solution can compete with on-premises in the case of this model company is to by applying data deduplication with a ration of at least 4:1.

If we look at the objectives, on-premises data storage does not meet durability objective in RAID 10 configuration with two disks in a mirror, although in RAID 6 configuration all objectives are met. On the other hand, Nearline Storage with its only 99% availability does not meet availability objective specified in SLA.

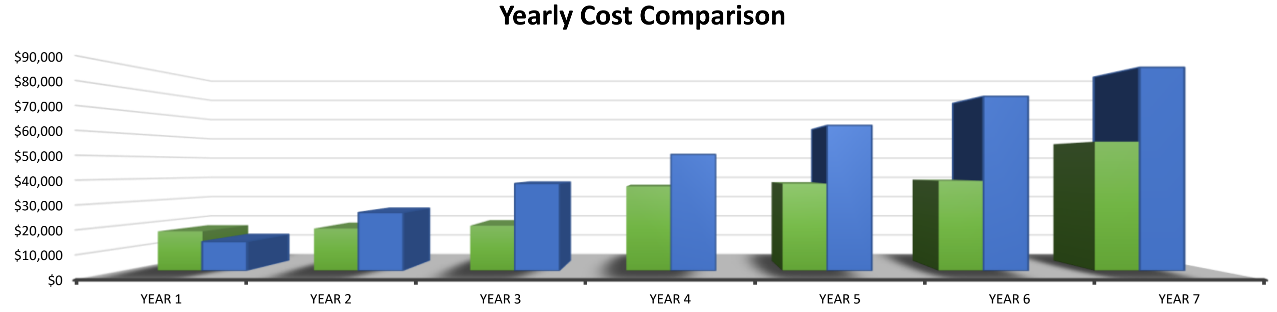
Figure 2.3: Model company 1: cost comparison of on-premises and cloud storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
🟩 On-premises	$10,816	$11,840	$12,864	$23,680	$24,704	$25,728	$36,544
🟦 Cloud	$6,658	$13,316	$19,974	$26,632	$33,290	$39,948	$46,606

A new comparison of all storage types that meet the objectives, along with the hybrid, is shown in the figure 2.4 and 2.5 As we can see, the hybrid solution is significantly cheaper than on-premises and cloud solution and offers more benefits at the forefront of meeting the backup recommendation 3-2-1 (see subsection 1.7).

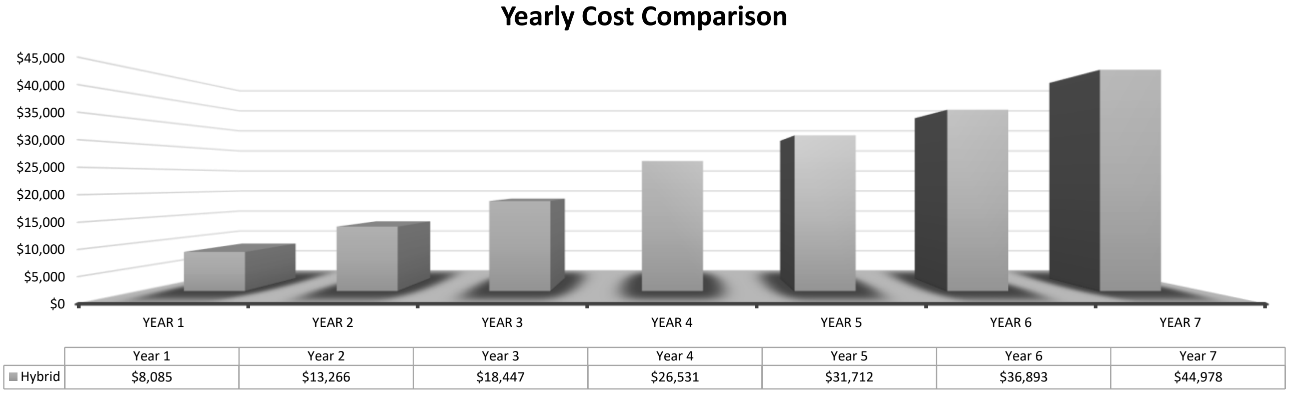
Figure 2.4: Model company 1: cost comparison of improved on-premises and cloud storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
🟩 On-premises	$17,155	$18,384	$19,613	$36,768	$37,997	$39,226	$56,381
🟦 Cloud	$12,697	$25,395	$38,092	$50,789	$63,487	$76,184	$88,881

The hybrid solution consists of DiskStation DS716+ with four 8 TB disks in RAID 10 configuration and Nearline storage. Backups from the first week

Figure 2.5: Model company 1: cost of hybrid storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
■ Hybrid	$8,085	$13,266	$18,447	$26,531	$31,712	$36,893	$44,978

are stored on-premises and the following three weeks of backups are stored in the Nearline storage.

In model company 1, if we want to save money even at the cost of not meeting the objective, we should stay with the on-premises solution. Otherwise, I recommend going into a hybrid solution that meets all objectives and backup recommendation 3-2-1.

Model Company 2

It is a medium-sized non-IT company with typical IT background (application servers, workstations, shared storages and databases) with the following information:

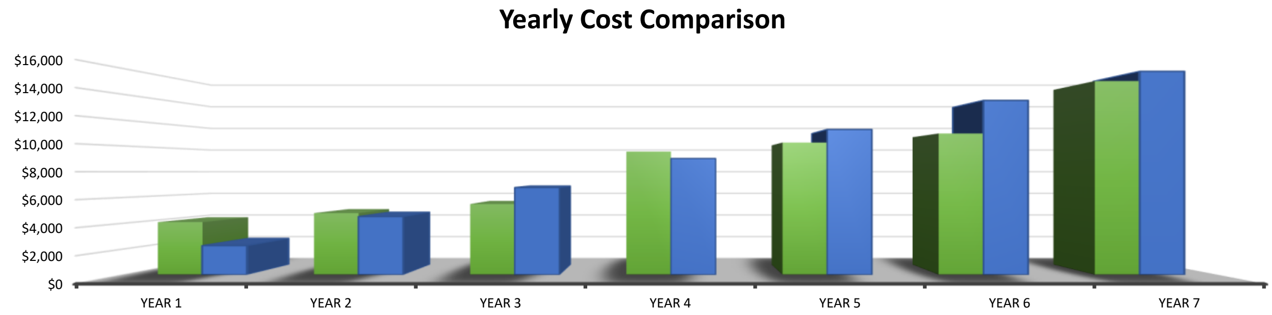
- Number of employees: 100
- Total size of the backed up data: 5 TB
- Internet connection speed: 250 Mb/s
- Data retention: 4 weeks
- RPO: 24 hours
- RTO: 36 hours
- Availability: 99.5%
- Durability: 99.99%

The total storage size was estimated to 18 TB, including compressed data, with no data deduplication, four incremental backups and one full backup a week with a ten percent data growth a year and three years of the refresh cycle.

For an on-premises storage server we selected NAS DiskStation DS1817+, with a 6 TB Enterprise Capacity 3.5 HDD from Seagate in RAID 6 configuration. As a cloud storage class we selected Nearline Storage.

When comparing the price of on-premises and cloud solution without retrieval cost the cost is comparable (see figure 2.6), however with retrieval cost the cloud solution is more expensive. On-premises meets all objectives. If we would want in cloud storage to meet the availability objective we would have to change storage class, which is again costly.Therefore on-premises is a more appropriate solution.

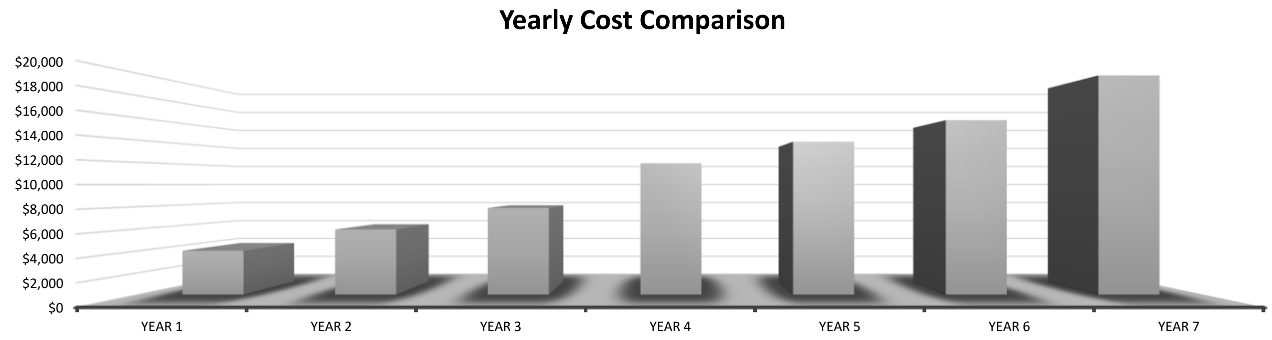
Figure 2.6: Model company 2: cost comparison of on-premises and cloud storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
🟩 On-premises	$4,078	$4,782	$5,485	$9,564	$10,267	$10,970	$15,049
🟦 Cloud	$2,256	$4,512	$6,769	$9,025	$11,281	$13,537	$15,794

A hybrid solution cost is shown in figure 2.7. As we can see the cost is higher than in case of on-premises, but it offers more benefits at the forefront of meeting the backup recommendation 3-2-1.

Figure 2.7: Model company 2: cost of hybrid storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
■ Hybrid	$4,023	$5,956	$7,890	$11,912	$13,846	$15,779	$19,802

The hybrid solution consists of DiskStation DS716+ with two 10 TB disks in RAID 1 configuration and nearline storage. Backups from the first week are stored on-premises and the following three weeks of backups are stored in the Nearline storage.

In model company 2, the most suitable solution is on-premises. However, if we want a more robust solution and we do not mind the higher price, we can choose a hybrid solution instead.

Model Company 3

It is a small e-commerce company. Everything is running on-site because e-shop is large and needs higher performance than shared web hosting offers. The system consists of an email client, a web server and a database. Further information about the company is as follows:

- Number of employees: 15
- Total size of the backed up data: 650 GB
- Internet connection speed: 100 Mb/s
- Data retention: 2 weeks
- RPO: 4 hours
- RTO: 24 hours
- Availability: 99.9%
- Durability: 99.99%

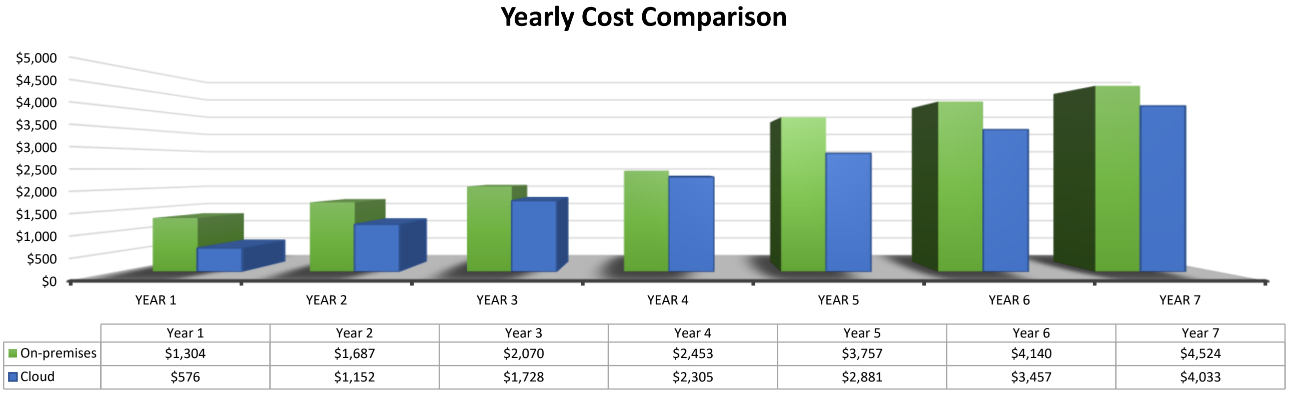
The total storage size was estimated to 2 TB, including compressed data, with no data deduplication, six incremental backups a day, four times a week and one full backup a week with a ten percent data growth a year and four years of the refresh cycle.

For an on-premises storage server we selected NAS DiskStation DS716+II, with a 2 TB IronWolf PRO from Seagate, due to insufficient durability with a 2 TB IronWolf, in RAID 1 configuration. As a cloud storage class we selected GCS Regional, due to the need for higher availability.

Both, on-premises and cloud storage, meet the objective. As we can see in the figure 2.8 in the current configuration the cloud storage is cheaper and outweigh on-premises benefits especially in terms of scalability and durability. The hybrid solution has not been compared, as the amount of data is small therefore it is clearly not worth investing in the local storage together with the cloud.

In model company 3 the cloud data backup storage is the winner in all the cost-benefit factors, so we should go to the cloud.

Figure 2.8: Model company 3: cost comparison of on-premises and cloud storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
🟩 On-premises	$1,304	$1,687	$2,070	$2,453	$3,757	$4,140	$4,524
🟦 Cloud	$576	$1,152	$1,728	$2,305	$2,881	$3,457	$4,033

Model Company 4

It is IT startup developing a web application. We assume that for some reason we do not want to use the cloud for all our data or applications so we have some data on-site needed to be backed up. Further information about the company is as follows:

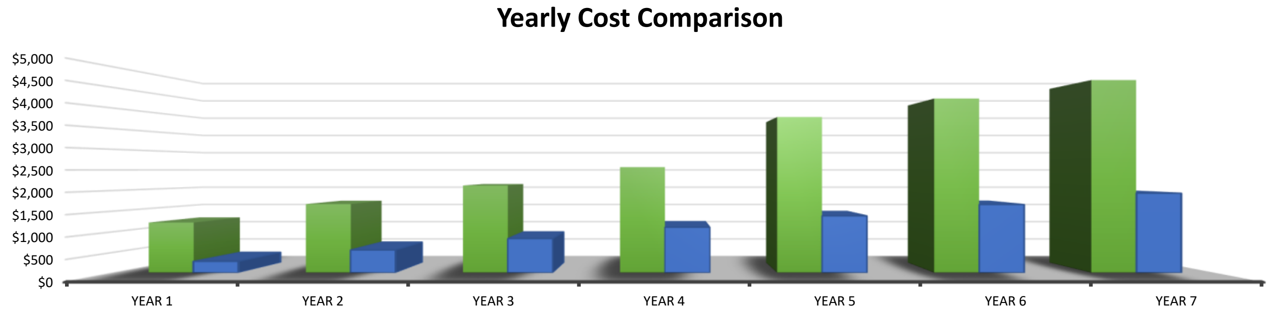
- Number of employees: 6
- Total size of the backed up data: 100 GB
- Internet connection speed: 100 Mb/s
- Data retention: 2 weeks
- RPO: 12 hours
- RTO: 24 hours
- Availability: 99%
- Durability: 99.99%

The total storage size was estimated to 200 GB, including compressed data, with no data deduplication, eight incremental backups and one full backup a week with a ten percent data growth a year and four years of the refresh cycle.

For an on-premises storage server we selected NAS DiskStation DS716+II, with a 1 TB IronWolf from Seagate, due to a disk compatibility with NAS, in RAID 1 configuration. As a cloud storage class we selected Nearline Storage.

The cloud solution, unlike the on-premises, meets all the objectives and it is much cheaper (see figure 2.9), due to poor on-premises scalability with such a small amount of data. The hybrid solution has not been considered, as the cloud solution itself is entirely sufficient.

Figure 2.9: Model company 4: cost comparison of on-premises and cloud storage.



	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
🟩 On-premises	$1,219	$1,669	$2,119	$2,569	$3,788	$4,238	$4,688
🟦 Cloud	$276	$552	$829	$1,105	$1,381	$1,657	$1,933

In model company 4 the cloud data backup storage is the winner in all the cost-benefit factors, so we should go to the cloud.

2.7 Results

Choice of backup solution depends on many factors and cannot be simply generalized. For a more accurate result, we should always make own analysis of our specific business environment (for that reason we can use the presented cost-benefit analysis tool in section 2.5). However, during the cost-benefit analysis in the previous section 2.6 we managed to make some general conclusions and recommendations.

Set of general recommendations for the backup in SMEs

Storing data in the cloud itself is cheaper than storing data on-premises. But if we add costly charges associated with moving data out of the cloud and optimization of data transfer, the cloud solution becomes more expensive. This problem can be solved by a hybrid solution.

In a small amount of data, despite the expensive fees associated with data retrieval, the cloud pays off, especially due to its scalability and durability. Furthermore, the benefits of an on-premises solution begin to prevail. And for a large amount of data, it is best to go into a hybrid solution, due to a 3-2-1 rule.

Set of general recommendations for the backup in SMEs based on the total data size stored in the storage is shown in the following table 2.2. The recommendations take into account only backup types that meet the objectives and are comparable among themselves in terms of cost and benefits. However, if hybrid solution is recommended, we should choose hybrid backup since it combines the benefits of both on-premises and cloud and meets the 3-2-1 rule.

Table 2.2: Set of general recommendations for backup in SMEs.

Total storage size	Recommendation	Explanation
< 2.5 TB	Cloud	It is cheap and it offers more benefits than on-premises (especially scalability and durability).
$\langle 2.5$ TB; 5 TB)	Cloud On-premises Hybrid	Decision depends on specific company and its objectives.[1]
$\langle 5$ TB; 10 TB)	On-premises Hybrid	Choose on-premises to save budget. Otherwise choose hybrid.
$\langle 10$ TB; 40 TB$\rangle$	On-premises	In a sufficient RAID 6 configuration it is cheaper than other solutions and it meets all objectives.
> 40 TB	Hybrid	Although the price is comparable to an on-premises, the hybrid offers more benefits.

[1] The factors most influencing the decision are: number of recovery per year, data backup objectives, storage class type, internet bandwidth and data deduplication.

2.8 Summary

In the previous chapter, a cost-benefit analysis of hybrid cloud backup for SMEs was performed. The analysis proceeded as follows. First, we defined the scope of the analysis. Second, we presented the cost and benefit factors of data backup and used them to design a mathematical model. Third, we created a tool for a calculation of the mathematical model. Then, we used the created tool for an analysis of four SMEs differing mainly according to the total size of the data to be backed up. Finally, we introduced a set of general recommendations for the backup in SMEs based on previous results.

Chapter 3

Data backup in the Czech Republic

In this chapter, a practical test of backup products available in the Czech Republic will be performed to verify the results of cost-benefit analysis from the previous chapter. Based on the results of testing, a new recommendation for backing up in SMEs in the Czech Republic will be presented.

First, the research of existing backup products that allow performing a backup of data to a cloud environment suitable for SMEs in the Czech Republic will be performed. Then, found backup products will be tested within available resources. Finally, based on the previous testing a set of new recommendations for the backup in SMEs in Czech Republic applying both technological and managerial dimension terms will be presented.

3.1 Research of existing products in the Czech Republic

Product requirements

When choosing a backup product, we need to ask ourselves what our expectations are or what our company needs from the backup product in order to meet business requirements. For instance, these may be the considered questions:

- Does the product support cloud integration?
- Does the product back up our operating systems, virtual machines or applications?
- What features does the product offer (e.g. synthetic backup, data deduplication, data compression, encryption, or automation)?

Product selection

There are not as many backup products available on the Czech market as abroad. The two best-known backup products here are Veeam and Acronis. This year, the Czech market expanded with Xopero.

The first representative of backup products is Veeam Backup Essentials 9.5. The product is designed to backup virtual machines. Veeam is known not only in the Czech Republic but also worldwide as a leader ([78] and [79]) in all segments of the market offering products for small, medium and large businesses with a wide range of features [80].

The second representative of backup products is Xopero Backup & Restore. This product is new, unexplored and offers most of the important features at a good price.

The third representative of backup products is Acronis Backup 12.5. This product offers similar features like the Veeam with the main differences in licensing, its cost and cloud support.

3.2 Testing of the products

Purpose of testing

The purpose of testing is to verify the advanced features of backup products, which are mainly used when using cloud backups with insufficient internet connection speed. Another reason is to verify the internet connection to the data center of public cloud providers. Due to the distance of the data centers, various connection problems may occur (latency, bandwidth throttling).

Test environment

On-site backup server:

- OS: MS Hyper-V server 2012 R2
- Guaranteed internet bandwidth: 30 Mb/s (symmetrical)
- RAM: 64 GB
- CPU: 4-Core
- Backup products:
  - Veeam Backup & Replication 9.5
  - Xopero Backup & Restore
  - Xopero Cloud
  - Acronis Backup 12.5

Cloud Services Providers:

- Azure (free trial)
- AWS (free trial)

Google is excluded from the list, because it does not offer free tier in Europe.

Veeam Backup & Replication 9.5

Hybrid deployment

Veeam does not support cloud backup natively [81]. If we want to deploy our own hybrid cloud backup, we will need additional appliances, therefore implementation is more complex and the cost of the overall solution including HW resources, deployment and maintenance is more expensive. For example, in the case of Amazon we will need AWS Storage Gateway and in the case of Azure, we can use StorSimple or Cloud Connect.

Due to the current lack of Veeam native support for Azure, Google and AWS, we should consider using of one of the Veeam partners (listed on the Veeam page [82]), who are offering the data backup service to their own data centers, which are located in the same country. The advantage is the simplicity of the solution and the fact that the data center is located in the same country. However, the disadvantage is that this service could be much more expensive than in the case of Azure, Google or AWS. Prices range from warm storage class to hot storage and more, but it is mostly a flat rate pricing for stored data, so we do not have to pay for a retrieving our data and other cloud usage fees.

We decided to use Veeam Cloud Connect as one of the ways to connect to Azure. The disadvantage of this solution is that we cannot use cheap Azure Blob storage. Hybrid deployment consists of on-premises Veeam Backup & Replication 9.5 and Veeam Cloud Connect for Service Providers running on the Azure's VM. A hybrid deployment and its individual steps are as follows:

Azure deployment:

1. New application: Veeam Cloud Connect for Service Providers
2. Connect to created VM and follow the guide
3. Insert Veeam Cloud Connect license
4. Add backup repository (location: West Europe)
5. Manage certificates
6. Add cloud gateway

7. Add tenant

On-premises deployment:

1. Install and verify Veeam Backup & Replication 9.5
2. Add service provider

For a more detailed description of deploying a hybrid solution, see blog post [83].

Test setup and results

Due to limited resources, only basic testing was performed, consisting of:

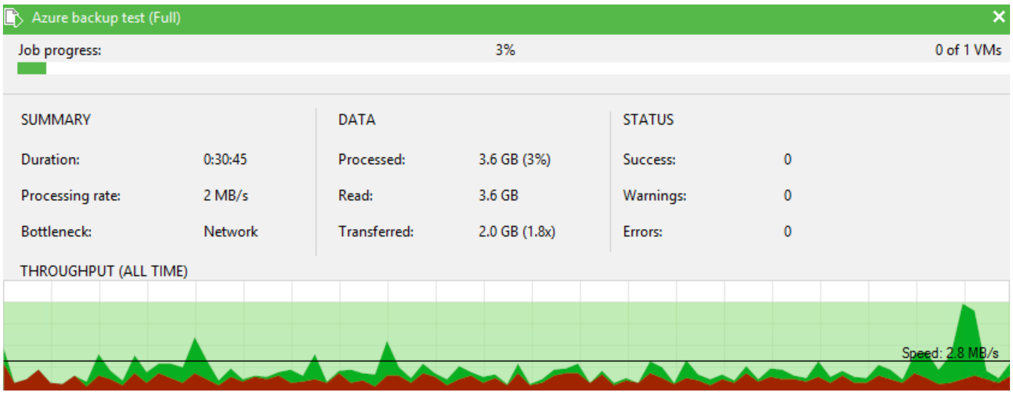
- Backup of one virtual machine
  - Locally
  - To the cloud
- Local backup of multiple virtual machines

During the local backup the size of the first VM was reduced by compression the storage space from 40.3 GB to 16.6 GB, which is 2.4 times less than the original size. The size of the second VM was reduced even more by 5.8 times. Other local backups had a compression ratio better than 2:1. The compression level was set to *Optimal (recommended)*.

In the figure 3.1 we can see the continuous data compression with the data deduplication achieving 1.8:1 ratio (Transferred), with the upload speed on average 8.6 Mb/s making the upload speed 15.6 Mb/s (Processing rate) of reduced original file after applying WAN optimization. That is half of our internet bandwidth.

When uploading a backup to the cloud, an additional cost factor was found. A backup job sent 3.15 GB to the Azure and 98 MB out of Azure (see figure 3.2). This can be converted to the fact that for every 32 GB transferred to the cloud, we will pay 1 GB of data coming out of the cloud. The data sent out of the cloud is the most expensive storage cost factor for cloud, and because the data is often sent to the cloud, this hidden cost factor can make a big difference in the final price of data backup storage solution. This cost was not verified due to a lack of network resources to transfer more data into the Azure.

Figure 3.1: Viewing real time backup to Azure statistics.



Azure backup test (Full)

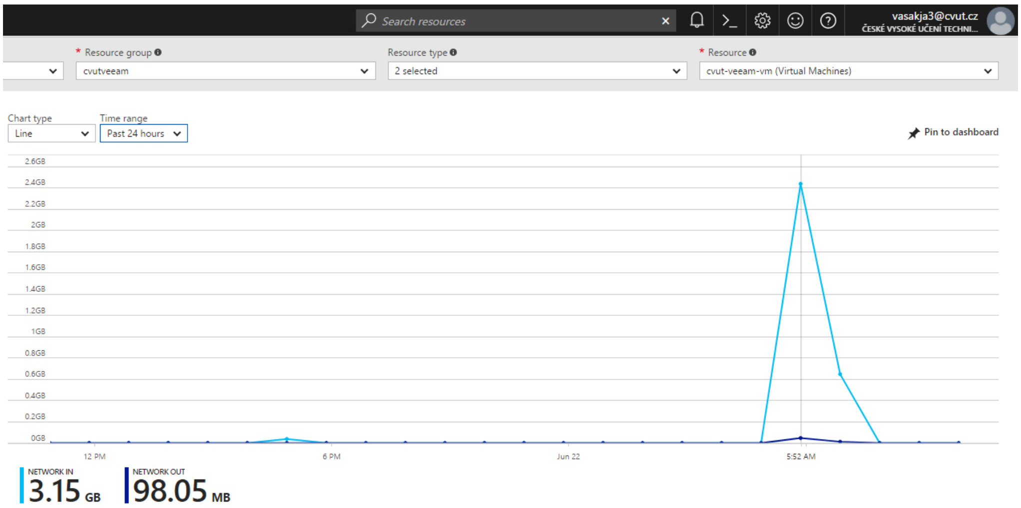
Job progress: 3% 0 of 1 VMs

SUMMARY		DATA		STATUS	
Duration:	0:30:45	Processed:	3.6 GB (3%)	Success:	0
Processing rate:	2 MB/s	Read:	3.6 GB	Warnings:	0
Bottleneck:	Network	Transferred:	2.0 GB (1.8x)	Errors:	0

THROUGHPUT (ALL TIME)

Speed: 2.8 MB/s

Figure 3.2: The amount of data sent out of Azure during data transfer to Azure.



Resource group: cvutveeam

Resource type: 2 selected

Resource: cvut-veeam-vm (Virtual Machines)

Chart type: Line

Time range: Past 24 hours

NETWORK IN 3.15 GB

NETWORK OUT 98.05 MB

Summary

Based on simple testing, the following information was found:

- Compression ratio is better than 2:1.
- Need for additional virtual appliance and costs associated with it (HW resources, deployment and maintenance).
- Costs that were not counted in the analysis:
  - The cost for public IP.

  - The cost for data transfer out hidden inside data transfer into the cloud.

- The upload speed to data center in the Netherlands is slow.

Xopero

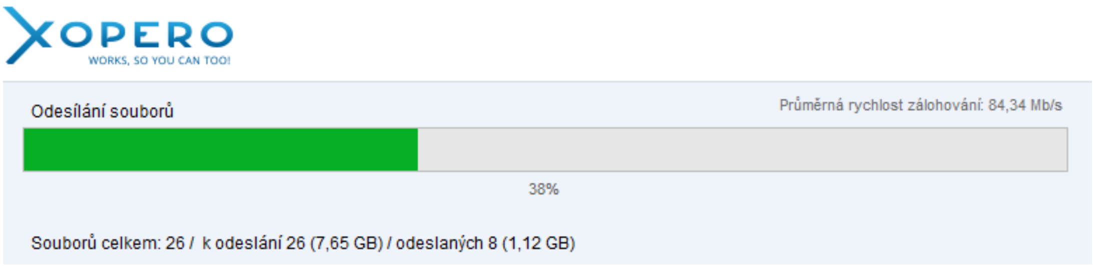
Xopero Backup & Restore

I did not have a positive experience with this product and I could not test it properly. I managed to deploy the product in several attempts. During the backup, the product reported a variety of errors and because the product is new on the market there is no information on the internet on how to deal with those issues yet. I assume that some errors are caused either by my ignorance about the product or its immaturity that will be resolved over time. I did not find some offered features, such as data deduplication. The product does not support the ability to connect to a cloud provider and thus the ability to create our own hybrid cloud. Despite these shortcomings, I still think this product can be an interesting and affordable option.

Xopero Cloud

The product is simple and clear. The backup to the cloud was fast. In an environment with an internet bandwidth 100 Mb/s the transfer speed into the cloud was 84 Mb/s (as we can see in the picture 3.3). The transfer speed out of the cloud was slower, with an average speed of only 32 Mb/s. The product lacks WAN optimization features as data compression, data deduplication or WAN acceleration. There is no specific SLA information on the product page.

Figure 3.3: Uploading data to the Xopero cloud.



Odesílání souborů

Průměrná rychlost zálohování: 84,34 Mb/s

38%

Souborů celkem: 26 / k odeslání 26 (7,65 GB) / odeslaných 8 (1,12 GB)

Conclusion

Xopero is better suited for smaller companies who want to save financial resources. I recommend using it in case we want to go to the cloud and we have enough internet bandwidth due to data volume.

Acronis Backup 12.5

Hybrid deployment

There are two ways on how to deploy a hybrid cloud within an Acronis Backup 12.5.

The first is to use the cloud storage offered directly by Acronis, called Acronis Cloud Storage. To use it, we only need to add a new cloud storage in the Acronis Backup 12.5 by adding our Acronis account. It cannot be easier, but the disadvantage is the price for 1 GB, which ranges from $0.049 to $0.099 per month.

The second is to use Acronis Storage Gateway, which supports Microsoft Azure Storage, Amazon S3 and others storage backends [84]. It runs on Linux distribution Red Hat Enterprise Linux (RHEL) or CentOs.

Due to the complexity of the second solution, I decided to deploy only a simple hybrid cloud with the Acronis Cloud Storage.

Test setup and result

Due to limited resources, only basic testing was performed consisting of:

- Local backup of one virtual machine.
- Local backup of multiple virtual machines.
- File/folder backup to the Acronis Cloud Storage.

Data compression of local backups was on average reduced by 3.2 times of the original data size. No data compression was performed when data was directly backed up to the cloud. The upload speed to the cloud was on average 17 Mb/s which is about half of the internet bandwidth.

3.3 Results

In simple terms, I recommend using Xopero for small businesses who want to save financial resources or go into the cloud, Veeam for SMEs using on-premises backups and Acronis for businesses that want to deploy their own hybrid backups. But since hybrid backups can be deployed by both products Veeam and Acronis, I would make a decision based on a cost analysis of total cost of hybrid data backup solution for our business which includes the product price and the price for a cloud storage or cloud provider.

Two important discoveries affecting the analysis of the previous chapter 2 were found when testing products.

The first discovery concerns performance. Data compression and data deduplication reaches a higher ratio than the 2:1. On the other hand, the throughput of service providers from a foreign country is 2-4 times lower than our internet bandwidth. But if we put together these two facts, we will get approximately the same bit rate as the one we used in the analysis.

The second discovery is related to the cost. New costs associated with cloud backup have been found. These are the cost of uploading data to the cloud, a public IP address (may be the case for an on-premises solution as well) and other appliances.

Another important finding of testing is that products do not usually support inexpensive backup storage from Google, AWS, or Azure, so we have to consider the higher price per GB offered by other service providers in a flat rate for data stored.

Set of recommendation for the backup in SMEs in the Czech Republic

The main conclusive point is that the price of the cloud solution will be higher than expected and the price of the hybrid data backup will be even higher because in our analysis we did not calculate with retrieving cost which is not included in flat rate pricing for storage data mostly offered by other service providers. Taking this into account, modified recommendations for the most suitable backup solution for SMEs in the Czech Republic is shown in the following table 3.1.

Table 3.1: Set of recommendations for backup in SMEs in Czech Republic.

Total storage size	Recommendation		Explanation
< 2.5 TB	Xopero Acronis	Cloud	Simplicity and scalability.
$\langle 2.5$ TB; 5 TB$)$	Veeam Acronis	On-premises Hybrid	Choose on-premises to save budget. Otherwise choose hybrid.
$\langle 5$ TB; 40 TB$\rangle$	Veeam	On-premises	In a sufficient RAID 6 configuration it is cheaper than other solutions and it meets all objectives.
> 40 TB	Acronis	Hybrid	Although the price is comparable to an on-premises, the hybrid offers more benefits. It is expected to use an inexpensive storage from Google or AWS.

3.4 Summary

In the previous chapter, testing of backup products found during a research of existing products in the Czech Republic was performed. Testing has verified the core functionality of the products (full and incremental backup of files, folders or virtual machines, backup to the cloud and automatic backup), along with some advanced techniques (data compression and data deduplication). It has been found that products available in the Czech Republic allow us to achieve company's objectives.

Furthermore, the new cost factors that were not included in the previous chapter were found. This is primarily the cost associated with appliances needed for backup to the cloud. These costs consist of more complex deployment, maintenance and the need for additional resources. It was also found that due to products lack of support for cloud providers (Google, AWS and Azure) and limited data transfer to a remote cloud data centers, in case we want to back up to the cloud, we will most likely have to use the services of more expensive local providers.

For this reason, the price difference between the local backup and the cloud backup can be so great that it is no longer worth using both cloud backups and hybrid backups in most of the cases, despite the clear benefits the solutions offers (off-site backup, scalability, durability and minimal maintenance).

Conclusion

The cost-benefit analysis from the second chapter confirms the hypothesis that cloud providers such as Azure, Google and AWS, offer a better price per GB than on-premises storages. However, if we add expensive cloud data recovery fees or other WAN optimization charges, cloud solutions become more expensive.

Since the first hypothesis is confirmed, and cloud providers offer a more cost-effective solution for 1 GB of data stored, the second hypothesis is then taken into consideration. The second hypothesis is using optimal hybrid backup solutions, both in terms of price and performance. This hypothesis was confirmed in most situations. The situation where the hybrid solution did not become a winner is when we do not have sufficient data to backup (the amount of capacity for backup storage required is less than 2.5 TB), so in this case, it is better to go into the cloud, or when we can use the RAID 6 configuration to meet the goals, which is better to remain within on-premises.

In the third chapter, we tested real backup products available in the Czech Republic to verify the theoretical statement from the second chapter. During the product testing, other price factors associated with cloud solutions were found, which were not anticipated in the previous analysis. These were mainly the costs associated with the need for additional appliances and the use of local cloud providers offering different prices than the providers mentioned in the analysis. Local cloud providers offer a flat rate per 1 GB, resulting in a more expensive backup solution. In this case, we reject the first hypothesis that the cloud solution is cheaper than the local solution, thus the hybrid solution is not the optimal solution anymore.

The suitability of the backup solution depends on our business requirements and goals. In general, we can say that if the backup product supports inexpensive cloud storages such as Amazon, Google and AWS, then it is preferable to use a hybrid cloud even at the cost of paying more than for local backup solutions because the hybrid cloud brings us more advantages.

All goals have been met. Cost and benefit factors were defined afterward, a mathematical model including the factors was proposed. The model can be used as a general assessment for the backup solution in terms of both cost and benefit. Furthermore, a tool to compare local and cloud backup solutions was developed. It can also be used to analyse hybrid solutions. In addition, the tool was used in the analysis of the model for small and medium-sized businesses. Finally, from the analysis came a set of general recommendations on where to back up. The general recommendations were verified by the practical testing of backup products which were found in the Czech Republic and a new modified set of recommendation was created.

During work, practical usage of inexpensive cloud storage from Google, Azure, or AWS were not verified. Work was left out of tape storage analysis, although it may be a suitable backup alternative for medium-sized businesses. Furthermore, the option to backup off-site on our personal device was omitted. Due to an insufficient internet connection, cloud backup was not tested more thoroughly. Also, we did not consider more advanced WAN optimization techniques or security issues.

This thesis can be used in the real world by all small and medium-sized companies looking for a suitable backup solution. The thesis presents a set of general recommendations supplemented by a specific set of recommendations for the Czech Republic. For more accurate results, we recommend that you conduct your own analysis to your business environment, as you can either use the suggested mathematical model or the created comparison tool.

In the future, we can go deeper into the parameters that affect data transfer, such as data compression, data deduplication, or WAN optimization. These parameters affect the duration of backups and affect the resources on the source node or target node. Another essential parameter is security. Encryption makes it difficult to connect to some cloud repositories or the use of data deduplication.

Bibliography

[1] Kashef, M. M.; Altmann, J. *A Cost Model for Hybrid Clouds*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ISBN 978-3-642-28675-9, pp. 46–60, doi:10.1007/978-3-642-28675-9_4. Available from: `http://dx.doi.org/10.1007/978-3-642-28675-9_4`

[2] Nanath, K.; Pillai, R. A Model for Cost-Benefit Analysis of Cloud Computing,. *Journal of International Technology and Information Management*, volume 22, no. 6, 2013. Available from: `http://scholarworks.lib.csusb.edu/jitim/vol22/iss3/6`

[3] Konstantinos, K.; Persefoni, M.; et al. Cloud Computing and Economic Growth. In *Proceedings of the 19th Panhellenic Conference on Informatics*, PCI '15, New York, NY, USA: ACM, 2015, ISBN 978-1-4503-3551-5, pp. 209–214, doi:10.1145/2801948.2802000. Available from: `http://doi.acm.org/10.1145/2801948.2802000`

[4] Chandra, D. G.; Borah, M. D. Cost benefit analysis of cloud computing in education. In *2012 International Conference on Computing, Communication and Applications*, Feb. 2012, ISSN 2325-6001, pp. 1–6, doi:10.1109/ICCCA.2012.6179142.

[5] Dutta, A. K.; Hasan, R. *How Much Does Storage Really Cost? Towards a Full Cost Accounting Model for Data Storage*. Cham: Springer International Publishing, 2013, ISBN 978-3-319-02414-1, pp. 29–43, doi:10.1007/978-3-319-02414-1_3. Available from: `http://dx.doi.org/10.1007/978-3-319-02414-1_3`

[6] Alhazmi, O. H.; Malaiya, Y. K. Evaluating disaster recovery plans using the cloud. In *2013 Proceedings Annual Reliability and Maintainability Symposium (RAMS)*, Jan. 2013, ISSN 0149-144X, pp. 1–6, doi:10.1109/RAMS.2013.6517700.

[7] SherWeb. Cloud servers Total Cost of Ownership Comparison Tool. 2014. Available from: http://info.sherweb.com/cloud-tco.html

[8] SNIA. SOLID STATE STORAGE TOTAL COST OF OWNERSHIP CALCULATOR. 2009. Available from: https://www.snia.org/forums/sssi/programs/TCOcalc

[9] Platform, G. C. Google Cloud Platform Pricing Calculator. 2017. Available from: https://cloud.google.com/products/calculator/

[10] web serice, A. AWS Total Cost of Ownership (TCO) Calculator. 2017. Available from: https://awstcocalculator.com/

[11] Azure, M. Total Cost of Ownership (TCO) Calculator. 2017. Available from: https://www.tco.microsoft.com/Home/Calculator

[12] dictionary, S. The 2016 SNIA dictionary. 2016. Available from: https://www.snia.org/sites/default/files/dictionary/SNIADictionaryv2016_1.pdf

[13] EMC. data archiving. Available from: https://www.emc.com/corporate/glossary/data-archiving.htm

[14] backup. Dictionary.com. Available from: http://www.dictionary.com/browse/backup

[15] Techopedia. data compression. Available from: https://www.techopedia.com/definition/884/data-compression

[16] Beal, V. data deduplication. Available from: http://www.webopedia.com/TERM/D/data_deduplication.html

[17] durability. Business Dictionary. Available from: http://www.businessdictionary.com/definition/durability.html

[18] Barr, J. New: Amazon S3 Reduced Redundancy Storage (RRS). 2010. Available from: https://aws.amazon.com/blogs/aws/new-amazon-s3-reduced-redundancy-storage-rrs/

[19] Techopedia. recovery point objective. Available from: https://www.techopedia.com/definition/1032/recovery-point-objective-rpo

[20] Techopedia. recovery time objective. Available from: https://www.techopedia.com/definition/24250/recovery-time-objective--rto

[21] Commission, E. small and medium-sized enterprises. Available from: http://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/sme

[22] Dilip C., N. *Inside Windows Storage: Server Storage Technologies for Windows 2000, Windows Server 2003 and Beyond*, chapter Backup and Restore Technologies for Windows. Addison-Wesley Professional, first edition, 07 2003, ISBN 978-0-321-12698-6. Available from: http://www.informit.com/articles/article.aspx?p=99985&seqNum=3

[23] Rao, U. H.; Nayak, U. *Data Backups and Cloud Computing.* Berkeley, CA: Apress, 2014, ISBN 978-1-4302-6383-8, pp. 263–288, doi:10.1007/978-1-4302-6383-8_13. Available from: http://dx.doi.org/10.1007/978-1-4302-6383-8_13

[24] full backup. Backup4all.com. 2012. Available from: http://www.backup4all.com/kb/backup-types-115.html

[25] incremental; differential backup. Encyclopedia.com. 2004. Available from: http://www.encyclopedia.com/computing/dictionaries-thesauruses-pictures-and-press-releases/full-backup

[26] Gregory, S. Mirror backup. 2012. Available from: http://typesofbackup.com/mirror-backup/

[27] Beal, V. synthetic backup. Available from: http://www.webopedia.com/TERM/S/synthetic_backup.html

[28] Gregory, S. Local backup. 2012. Available from: http://typesofbackup.com/local-backup/

[29] off-site backup. YourDictionar.com. Available from: http://www.yourdictionary.com/off-site-backup

[30] TechNet. Backup and restore: classification of scenarios. 2010. Available from: https://social.technet.microsoft.com/wiki/contents/articles/206.backup-and-restore-classification-of-scenarios.aspx

[31] Rouse, M. What is image-based backup? 2011. Available from: http://searchdatabackup.techtarget.com/definition/image-based-backup

[32] Fourge, N. What is the difference between file backup and image backup? 2014. Available from: http://novabackup.novastor.com/blog/difference-between-file-backup-and-image-backup/

[33] Rouse, M.; Sullivan, E. cloud backup (online backup). 2016. Available from: http://searchdatabackup.techtarget.com/definition/cloud-backup

[34] Crump, G. What is hybrid cloud backup? 2014. Available from: https://storageswiss.com/2014/12/16/what-is-hybrid-cloud-backup/

[35] Dubey, S.; Sistla, V. T.; et al. *Microsoft System Center Data Protection for the Hybrid Cloud*. Pearson Technology Group, 2015. Available from: http://www.ebook.de/de/product/24329666/shreesh_dubey_vijay_tandra_sistla_shivam_garg_microsoft_system_center_data_protection_for_the_hybrid_cloud.html

[36] IDC. The digital universe of opportunities. 2014. Available from: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

[37] Venish, A.; Sankar, K. S. Study of Chunking Algorithm in Data Deduplication. In *Proceedings of the International Conference on Soft Computing Systems*, Springer Nature, dec 2015, pp. 13–20, doi:10.1007/978-81-322-2674-1_2.

[38] Meyer, D. T.; Bolosky, W. J. A Study of Practical Deduplication. In *Proceedings of the 9th USENIX Conference on File and Stroage Technologies*, FAST'11, Berkeley, CA, USA: USENIX Association, 2011, ISBN 978-1-931971-82-9, pp. 1–1. Available from: http://dl.acm.org/citation.cfm?id=1960475.1960476

[39] Rouse, M. global data deduplication. 2016. Available from: http://searchdatabackup.techtarget.com/definition/global-data-deduplication

[40] Acronis. Deduplication in Acronis Backup Advanced. 2016. Available from: www.acronis.com/en-us/download/docs/ABD/technical-whitepaper/

[41] Wainwright, A. WAN Optimization: What is it and what are the benefits? Available from: http://www.securedgenetworks.com/blog/WAN-Optimization-What-is-it-and-what-are-the-benefits

[42] RENO, N. Microsoft, Google and IBM Public Cloud Surge is at Expense of Smaller Providers. 2017. Available from: https://www.srgresearch.com/articles/microsoft-google-and-ibm-charge-public-cloud-expense-smaller-providers

[43] web services, A. Amazon S3 Pricing. 2017. Available from: https://aws.amazon.com/s3/pricing/

[44] web services, A. Amazon S3 Service Level Agreement. 2015. Available from: https://aws.amazon.com/s3/sla/

[45] web services, A. Amazon Glacier Pricing. 2017. Available from: https://aws.amazon.com/glacier/pricing/

[46] cloud platform, G. Overview of Storage Classes. 2017. Available from: https://cloud.google.com/storage/docs/storage-classes

[47] cloud platform, G. Google Cloud Storage SLA. Available from: https://cloud.google.com/storage/sla

[48] Azure, M. Azure Storage Pricing. 2017. Available from: https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

[49] Azure, M. SLA for Storage. 2016. Available from: https://azure.microsoft.com/en-us/support/legal/sla/storage/v1_0/

[50] Seagate. Solid-State Storage. Available from: https://www.seagate.com/files/www-content/product-content/pulsar-fam/_cross-product/en-gb/docs/ssd-faq-tp612-1-1003gb.pdf

[51] logic, S. Options in Tape Technology: TS1150 Technology and LTO. 2015. Available from: https://edge.spectralogic.com/index.cfm?fuseaction=home.displayFile&DocID=4732

[52] Sony. Optical Disc Archive Generation 2. 2016. Available from: http://www.everspan.com/wp-content/uploads/2016/04/ODA_Gen2_WhitePaper_20160408_English.pdf

[53] Leventhal, A. Triple-Parity RAID and Beyond. *Queue*, volume 7, no. 11, Dec. 2009: pp. 30:30–30:39, ISSN 1542-7730, doi: 10.1145/1661785.1670144. Available from: http://doi.acm.org/10.1145/1661785.1670144

[54] Harris, R. Why RAID 6 stops working in 2019. 2012. Available from: http://www.zdnet.com/article/why-raid-6-stops-working-in-2019/

[55] IBM. Re-Evaluating RAID-5 and RAID-6 for slower larger drives. 2016. Available from: https://www.ibm.com/developerworks/community/blogs/InsideSystemStorage/entry/Re_Evaluating_RAID_5_and_RAID_6_for_slower_larger_drives

[56] Moon, S.; Reddy, A. L. N. Don't Let RAID Raid the Lifetime of Your SSD Array. In *Proceedings of the 5th USENIX Conference on Hot Topics in Storage and File Systems*, HotStorage'13, Berkeley, CA, USA: USENIX Association, 2013, pp. 7–7. Available from: http://dl.acm.org/citation.cfm?id=2534861.2534868

[57] Schroeder, B.; Lagisetty, R.; et al. Flash Reliability in Production: The Expected and the Unexpected. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*, Santa Clara, CA: USENIX Association, 2016, ISBN 978-1-931971-28-7, pp. 67–80. Available from: `http://usenix.org/conference/fast16/technical-sessions/presentation/schroeder`

[58] Intel; Amplidata. Revolutionary Methods to Handle Data Durability Challenges for Big Data. 2012. Available from: `http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-amplidata-storage-paper.pdf`

[59] Shenoy, A. The Pros and Cons of Erasure Coding and Replication vs. RAID in Bext-Gen Storage Platforms. 2015. Available from: `http://www.snia.org/sites/default/files/SDC15_presentations/datacenter_infra/Shenoy_The_Pros_and_Cons_of_Erasure_v3-rev.pdf`

[60] Schroeder, B.; Gibson, G. A. Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? In *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, FAST '07, Berkeley, CA, USA: USENIX Association, 2007. Available from: `http://dl.acm.org/citation.cfm?id=1267903.1267904`

[61] Beach, B. How long do disk drives last? 2013. Available from: `https://www.backblaze.com/blog/how-long-do-disk-drives-last/`

[62] Sutter, D. J. How did Google lose, and find, all those e-mails? 2011. Available from: `http://edition.cnn.com/2011/TECH/web/03/01/gmail.lost.found/`

[63] Miller, R. Inside Facebook's Blu-Ray Cold Storage Data Center. 2015. Available from: `http://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/`

[64] Fruehe, J. Everspan changes archiving. Technical report, Moor Insights & Strategy, 2016. Available from: `http://www.moorinsightsstrategy.com/wp-content/uploads/2016/03/Everspan-Changes-Archiving-by-Moor-Insights-and-Strategy.pdf`

[65] EventHelix. System Reliability and Availability. Available from: `http://www.eventhelix.com/RealtimeMantra/FaultHandling/system_reliability_availability.htm`

[66] Pinheiro, E.; Weber, W.-D.; et al. Failure Trends in a Large Disk Drive Population. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, FAST '07, Berkeley, CA, USA:

USENIX Association, 2007, pp. 2–2. Available from: http://dl.acm.org/citation.cfm?id=1267903.1267905

[67] Klein, A. One Billion Drive Hours and Counting: Q1 2016 Hard Drive Stats. 2016. Available from: https://www.backblaze.com/blog/hard-drive-reliability-stats-q1-2016/

[68] Oggerino, C. *High Availability Network Fundamentals: A Practical Guide to Predicting Network Availability.* Cisco Press, first edition, 2001, ISBN 1587130173.

[69] Schwede, C. Swift object durability calculator. 2015. Available from: http://redhat-cip.github.io/swift-durability-calculator/

[70] Krogh, P. Backup Overview. 2015. Available from: http://www.dpbestflow.org/backup/backup-overview

[71] Floyer, D. Private cloud is more cost effective than public cloud for organizations over $1B. 2010. Available from: http://wikibon.org/wiki/v/Private_Cloud_is_more_Cost_Effective_than_Public_Cloud_for_Organizations_over_$1B

[72] Butler, B. Is there a point where a private cloud is cheaper than the public cloud. 2014. Available from: http://www.networkworld.com/article/2825994/cloud-computing/is-there-a-point-where-a-private-cloud-is-cheaper-than-the-public-cloud.html

[73] Microsoft. The economics of the cloud. 2010. Available from: https://news.microsoft.com/download/archived/presskits/cloud/docs/The-Economics-of-the-Cloud.pdf

[74] Harris, R. Availability vs durability in archive solutions. 2016. Available from: http://www.hgst.com/sites/default/files/resources/StorageMojo-Availability-vs-durability.pdf

[75] Ponemon. 2013 Cost of Data Center Outages. resreport, Ponemon institute, 2013. Available from: http://www.ponemon.org/library/2013-cost-of-data-center-outages?s=2013+Cost+of+Data+Center+Outages

[76] Young, E. . Data loss prevention. 2011. Available from: http://www.ey.com/publication/vwluassets/ey_data_loss_prevention/$file/ey_data_loss_prevention.pdf

[77] Ponemon. 2016 Cost of Data Breach Study: Global Analysis. resreport, Ponemon Institute, 2016. Available from: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=SEL03094WWEN

[78] Crowd, G. Best Backup Software. Available from: https://www.g2crowd.com/categories/backup?segment=all

[79] Russell, D.; Rinnen, P.; et al. Magic Quadrant for Data Center Backup and Recovery Software. 2016. Available from: https://www.gartner.com/doc/reprints?id=1-38Z8E64&ct=160607&st=sb

[80] Veeam. The #1 VM Backup for small businesses. 2017. Available from: https://www.veeam.com/veeam_essentials_9_5_datasheet_ds.pdf

[81] Eremin, V. Azure BLOB in Veeam Backup & Replication. 2016. Available from: https://forums.veeam.com/veeam-backup-replication-f2/azure-blob-in-veeam-backup-replication-t30175.html

[82] Veeam. Find Veeam-powered services. Available from: https://www.veeam.com/find-a-veeam-cloud-provider.html

[83] Serre, R. Make a Veeam backup copy to Microsoft Azure. 2017. Available from: https://www.tech-coffee.net/make-a-veeam-backup-copy-to-microsoft-azure/

[84] Acronis. Acronis Storage Gateway. 2017. Available from: dl2.acronis.com/u/pdf/acronis_storage_gateway_deployment_en-US.pdf

Appendix A

Acronyms

AFR Annualized failure rate

AWS Amazon Web Services

CRR Cross Region Replication

CSP Cloud Service Provider

DR Disaster Recovery

GCS Google Cloud Storage

GRS Geo Redundant Storage

HDD Hard Disk Drive

HW Hardware

IA Infrequent Access

IT Information Technology

LRS Locally Redundant Storage

MTBF Mean Time Between Failures

MTTR Mean Time To Repair

NAS Network-Attached storage

OS Operating System

RA-GRS Read Access - Geo Redundant Storage

RAID Redundant Array of Independent Disks

RHEL Red Hat Enterprise Linux

RFR Rebuild Failure Rate

RPO Recovery Point Objective

RTO Recovery Time Objective

R/W Read/Write

SME Small and Medium-sized Enterprise

SD Secure Digital

SDD Solid-State Drive

SLA Service Level Agreement

SW Software

TCO Total Cost of Ownership

UPS Uninterruptible power supply

URE Unrecoverable Read Error

USB Universal Serial Bus

WAN Wide Area Network

WORM Write Once Read Many

ZRS Zone Redundant Storage

Appendix B

Used Hardware for analysis

B.1 Used disks

For analysis in 2 chapter were used disks from Seagate:

- Seagate Enterprise Capacity 3.5 HHD (see the following table B.1)
- Seagate IronWolf Pro NAS (see the following table B.2)
- Seagate IronWolf NAS (see the following table B.3)

Because of lower cost and better specification over disks from another competitor such as HGST (formerly Hitachi Global Storage Technologies) and WD (Western Digital). Where usage of Seagate Enterprise Capacity HDDs is recommended in environments with more than 5 disks.

Table B.1: Seagate enterprise Capacity 3.5 HDD specification overview.

Specifications	10 TB (Helium)	8 TB	6 TB	4 TB	2 TB
Cost	$420	$378	$292	$197	$135
Max. Sustained Transfer Rate	249 MB/s	249 MB/s	226 MB/s	226 MB/s	226 MB/s
AFR	0.35%	0.44%	0.44%	0.44%	0.44%
MTBF	2.5M	2M	2M	2M	2M
URE	1 per 10E15				
Power consumption					
• Idle	4.5W	7.6W	7.25W	5.45W	4.25W
• Operating	8W	11W	8.31W	6.94W	5.9W
Warranty	5 years				

Table B.2: Seagate IronWolf Pro NAS specification overview.

Specifications	10 TB	8 TB	6 TB	4 TB	2 TB
Cost	$415	$350	$270	$210	$145
Max. Sustained Transfer Rate	214 MB/s	214 MB/s	214 MB/s	214 MB/s	195 MB/s
AFR			0.73%		
MTBF			1.2M		
URE			1 per 10E15		
Power consumption					
• Idle	4.4W	4.4W	7.6W	6W	4.5W
• Operating	6.8W	6.8W	9W	6.7W	7.5W
Warranty			5 years		

Table B.3: Seagate IronWolf NAS specification overview.

Specifications	8 TB	6 TB	4 TB	2 TB	1 TB
Cost	$300	$245	$130	$80	$65
Max. Sustained Transfer Rate	210 MB/s	195 MB/s	180 MB/s	180 MB/s	180 MB/s
AFR			0.87%		
MTBF			1M		
URE	1 per 10E15			1 per 10E14	
Power consumption					
• Idle	4.42W	7.2W	3.95W	3.7W	2.5W
• Operating	6.8W	9W	4.8W	5W	3.6W
Warranty			3 years		

B.2 Used NAS

NAS from Synology (see the following table B.4) were selected over the competitor's and Do It Yourself (DIY) solutions due to the high recommendation from the users in forums because of setup, maintenance and purchase simplicity. The downsides are the higher cost and lower throughout. Synology NAS can be expanded by the expansion units shown in the following table B.6.

Table B.4: Synology NAS for SMEs specification overview.

NAS	DiskStation DS716+II	DiskStation DS916+	DiskStation DS1817+	DiskStation DS2415+
Price	\$449	\$599	\$1,015	\$1,339
Drive Bays	2 (up to 7)	4 (up to 9)	8 (up to 18)	12 (up to 24)
Scale up	1xDX513	1xDX513	2xDX513	1xDX1215
Memory	2 GB	8 GB	8 GB (up to 16 GB)	2 GB (up to 6 GB)
Lan Port[1]	2x1GbE	2x1GbE	4x1GbE	4x1GbE
Sequential throughput[2]				
• Read	226 MB/s	226 MB/s	450 MB/s[3]	450 MB/s
• Write	190 MB/s	222 MB/s	400 MB/s[3]	396 MB/s
Power Consumption[4]				
• Access	19W	30W	61W	73W
• Hibernation	9.50W	13W	32W	37W

[1] With Link Aggregation.

[2] Measured in the RAID5 configuration.

[3] The sequential throughput in 10GbE environment was measured 1,179 MB/s for read and 542 MB/s for write.

[4] Power consumption is measured when fully loaded with Western Digital 1TB WD10EFRX HDDs with power requirements:

- Operating 3.3W
- Idle 2.3W

Table B.5: Synology NAS for Large Scale Business specification overview.

NAS	DiskStation RS3617xs	DiskStation RS4017xs+	DiskStation RS18017xs+
Price	$2,600	$5,300	$6,000
Drive Bays	12 (up to 36)	16 (up to 40)	12 (up to 180)
Scale up	2xRX1217	2xRX1217	96 (RX1217sas) 180 (RX2417sas)
Memory	4 GB	8 GB	16 GB (up to 128 GB)
Lan Port	4x1GbE	4x1GbE 2x10GbE	4x1GbE 2x10GbE
Sequential throughput			
• Read	3,012 MB/s	5,960 MB/s	4,872 MB/s
• Write	1,635 MB/s	2,494 MB/s	2,078 MB/s
Power Consumption			
• Access	116W	136W	142W
• Hibernation	60W	69W	77W

Table B.6: Synology NAS expansion units specification overview.

Expansion unit	Price	Drive Bays
DX513	$489	5
DX1215	$1,124	12
RX1217	$1,375	12
RX1217sas	$2,500	12
RX2417sas	$3,000	24

Appendix C

Contents of enclosed CD

```
 readme.txt ....................... the file with CD contents description
 analysis.....................the directory with analysis of model SME
 src.....................................the directory of source codes
   thesis..............the directory of LaTeX source codes of the thesis
     figures.................the directory of figures used in the thesis
 text .......................................... the thesis text directory
   DP_Vasak_Jaroslav_2017.pdf ......... the thesis text in PDF format
 tool...................the directory with data backup comparison tool
   Data_Backup_Comparison_Tool.xlsx
```