Universidade de Brasília
Institute of Exact Sciences
Department of Computer Science

A Backup-as-a-Service (BaaS) Software Solution

Heitor M. de Faria

Dissertation presented as partial requirement for conclusion on the Professional Master in Applied Computing

Advisor
Prof. Dra. Priscila Solis

Brasília
2018

Universidade de Brasília
Institute of Exact Sciences
Department of Computer Science

A Backup-as-a-Service (BaaS) Software Solution

Heitor M. de Faria

Dissertation resented as partial requirement for conclusion do
Professional Master in Applied Computing

Prof. Dra. Priscila Solis (Advisor)
CIC/UnB

Prof. Dr. Jacir Bordim
Universidade de Brasília

Dr. Georges Amvame-Nzê
Universidade de Brasília

Prof. Dr. Marcelo Ladeira
Coordinator of the Post-graduation Program in Applied Computing

Brasília, July 1st, 2018

Abstract

Backup is a replica of any data that can be used to restore its original form. However, the total amount of digital data created worldwide more than doubles every two years and is expected reach 44 trillions of gigabytes in 2020, bringing constant new challenges to backup processes. Enterprise backup is one of the oldest and most performed tasks by infrastructure and operations professionals. Still, most backup systems have been designed and optimized for outdated environments and use cases. That fact, generates frustration over currently backup challenges and leads to a greater willingness to modernize and to consider new technologies. Traditional backup and archive solutions are no longer able to meet users current needs. The ideal modern currently backup and recovery software product should not only provide features to attend a traditional data center, but also allow the integration and exploration of the growing Cloud, including "backup client as a service" and "backup storage as a service". The present study aims to propose and deploy a Backup as a Service software solution. To achieve that, the cloud/backup parameters, cloud backup challenges, researched architectures and BaaS system requirements are determined. Then, we select a set of BaaS desired features to be developed, that results in the first truly cloud REST API based Backup-as-a-Service interface, namely "bcloud". We conduct an on-line usability inquiry with a significant number of users and perform a result analysis. The overall average objective zero to ten questions evaluation was 8.29%, indicating a very satisfactory user perception of the bcloud BaaS interface prototype.

Keywords: backup as a service - BaaS, cloud, disaster recovery

Contents

List of Figures

List of Tables

Acronyms

ABC Activity Based Costing.

AES Advanced Encryption Standard.

ANSI American National Standards Institute.

API Application Programming Interface.

BaaS Backup as a Service.

CDP Continuous Data Protection.

CLI Command Line Interface.

CPU Central Processing Unity.

CRAM Challenge-Response Authentication.

CSP Cloud Service Provider.

CSS Cascading Style Sheets.

DHT Distributed Hash Table.

DR Disaster Recovery.

EFI Extensible Firmware Interface.

EFS Encrypting File System.

FB Frequency of Backup.

FIFO First In First Out.

GRB Geographical Redundancy and Backup.

GUI Graphical User Interface.

HTTPS Hyper Text Transfer Protocol Secure.

IaaS Infrastructure as a Service.

IP Internet Protocol.

IPCS Inter-Private Cloud Storage.

IT Information Technology.

LAN Local Area Network.

LBS Local Backup Server.

LTO Linear Tape-Open.

LVM Logical Volume Management.

MD5 Message-Digest Algorithm 5.

NAS Network Attached Storage.

NAT Network Address Translation.

NDMP Network Data Management Protocol.

NIST National Institute of Standards and Technology.

OS Operating System.

PaaS Platform as a Service.

PC Personal Computer.

PHP PHP: Hypertext Preprocessor.

QoS Quality of Service.

RBS Remote Backup Server.

REST Representational State Transfer.

RPO Recovery Point Objective.

RTO Recovery Time Objective.

S3 Simple Storage Service.

SaaS Software as a Service.

SAN Storage Area Network.

SD2SD Storage Daemon to Storage Daemon.

SDDB Secure-Distributed Data Backup.

SHA1 Secure Hash Algorithm 1.

SLA Service Level Agreement.

SME Small and Medium Enterprises.

SNIA Storage Networking Industry Association.

SOA Service Oriented Architecture.

SQL Structured Query Language.

SSL Secure Sockets Layer.

TCP Transmission Control Protocol.

TLS Transport Layer Security.

UEFI Unified Extensible Firmware Interface.

UX User Experience.

VM Virtual Machine.

VSS Volume Shadow Copy Service.

VTL Virtual Tape Library.

XaaS Anything as a Service.

Chapter 1

Introduction

In the words of Guise [7], backup is the replica of any data that can be used to restore its original form. In other words, backup is a valid copy of data, files, applications or operating systems that can serve for the recovery purpose. That is why, customarily, backup is often stored to lower cost of high capacity removable media, such as magnetic tapes, that are stored in fireproof safes or another protected physical environment.

Conforming to the International Data Corporation (IDC) [8], the total amount of digital data created worldwide more than doubles every two years. It will grow from 4.4 zettabytes in 2013 to 44 zettabytes by 2020. Just for illustration, one zettabyte is equal to 1 trillion gigabytes. This astounding growth comes from both the number of devices generating data as well as the number of sensors in each device, and 80 percent of it is stored at enterprise data centers at some point, requiring backup and other security mechanisms for information maintenance.

According to Russell et al. [3], Russel [9], enterprise backup is among the oldest most-performed tasks for infrastructure and operations professionals. Still, most backup systems have been designed and optimized for outdated environments and use cases. That fact, generates frustration over currently backup challenges and leads to a greater willingness to modernize and to consider new technologies. As stated by Silva [10], traditional backup and archive solutions are no longer able to meet users current needs.

In line with Kaiser et al. [11], the massive increase in data generation and processing in industry and academia has dramatically increased the pressure on backup environments. Data itself has become a valuable asset, and it is more important to protect it using highly reliable backup systems. Nonetheless, most backup systems have been based on tape environments until the late 2000s, and magnetic tapes were still one of the most cost-effective ways to store data.

In the vision of Russell et al. [3], the ideal modern currently backup and recovery software products should not only provide features to attend a traditional data center, e.g.: backup to conventional random-access or sequential media (hard disk, solid-state, tape drives), data reduction techniques (compression, deduplication or single instancing) and systems interoperability. However, must also allow the integration and exploration of the growing Cloud, including “backup client as a service” and “backup storage as a service”.

Moreover, even the currently sold solutions such as Backup as a Service (BaaS) do not meet the new market requirements for data center backup and recovery software [3]. Backup software for a homogeneous environment is also excluded, such as native tools from Microsoft or VMware for their specific platforms. The majority of midsize and large customers prefer a single, scalable backup product for their entire environment.

According to Khoshkholghi et al. [12], rapid cloud computing development is motivating more industries to use a variety of cloud services. Many security challenges have been raised, such as risk management, trust and recovery mechanisms should be studied to provide business continuity and better user satisfaction.

1.1 Aim and Objectives

The present study aims to deploy a Backup as a Service (BaaS) solution, under the Remote Backup to the Cloud architecture. To achieve that, we determined the cloud/backup parameters, cloud backup challenges, researched architectures and BaaS system requirements. Then, a set of features are selected to be developed and implemented to attend the chosen architecture. We apply the user-inquiry technique with a significant amount of users to validate the developed features and, finally, we present the results and analysis.

1.2 Structure

This dissertation is structured as follows. Chapter 2 presents the Literature Review, the State-of-the-Art, concepts, requirements and related surveys. Chapter 3 describes the Bacula backup system which will be the basis for the development of a BaaS solution. Chapter 4, details the proposal, the development scope, execution, solution topology, BaaS application components, the usability evaluation, the user-inquiry minutiae, and results. Finally, Chapter 5 draw some conclusions and the project’s future works.

1.3 Related Author Publications

The following author studies provide some of the base for this dissertation:

- Bacula The Open Source Backup Software (book) [13].
- Storage Growing Forecast with Bacula Backup Software Catalog Data Mining [14].
- Backup Storage Block Level Deduplication with DDUMBFS and BACULA [15].
- A Hadoop Open Source Backup Solution [16].

Chapter 2

Literature Review

In this Chapter, we introduce the Cloud Computing and the Disaster Recovery Parameters. We list the Cloud-Based Disaster Recovery Challenges and present a Survey of Cloud Backup Architectures. Then, we exhibit an examination of the BaaS macro-requirements in comparison with current traditional distributed backup software features.

2.1 Cloud Computing

In correspondence with Armbrust et al. [17], Buyya et al. [18], cloud computing is a long-held idea of computing as a general utility. It promises to shift data and computational services from individual devices to distributed architectures. The initial cloud content was created to describe sets of complex on-demand services offered by commercial providers. Then, based on the advancement in Internet topology with larger bandwidth and mobile access dissemination, any individual can upload personal information to the cloud.

Cloud computing comprehends *Internet-based distributed computing platforms which are highly scalable and flexible* [12]. It is the allocation of IT resources (v.g.: computational power, storage, software, hardware platforms, and applications) to a wide range of consumers with different access devices.

As reported by Columbus [19], Ried et al. [20], Arean [21], cloud computing services global total market revenues value is expected to grow from U$ 67 billion by 2015 to U$ 241 billion by the end of 2020. Still in 2013, near to 61% of United Kingdom businesses were relying on some cloud services. Cloud computing is becoming more popular [12] in large-scale computing because of its ability to share globally distributed resources. The users can access cloud-based services anywhere in the world. The biggest Cloud Service Providers are developing data centers in several continents to support different cloud services.

In agreement with the National Institute of Standards and Technology (NIST) Definition of Cloud Computing [22], a cloud model is composed of five essential characteristics:

On-demand self-service. Meaning a consumer can unilaterally provision computing capabilities as needed automatically (e.g., time and network storage) without requiring human interaction with representatives from the service provider.

Measured service. Automatic control and resource optimization cloud systems shall have, by leveraging a metering capability at some level of abstraction, to appropriate services usages such as storage, processing, bandwidth, and active user accounts. Resource usage must be able to be monitored, reported and controlled, providing reliable information for both customer and service provider.

Resource pooling. Where the cloud provider's computing resources are pooled to serve multiple consumers under a multi-tenant model. The cloud dynamically assigns different physical and virtual resources according to demand. The customer generally has no control or knowledge over the exact location of the provided resources but might be able to specify a location at with a higher level of abstraction, such a country, state, or datacenter. Examples of resources pooled resources include processing, storage, memory and network bandwidth.

Rapid elasticity. It is the elastic provision of capabilities that can be released to scale rapidly outward or inward commensurate with demand, in many cases automatically. From the consumer point of view, the capabilities available for provisioning should appear to be unlimited and can be appropriated at any time and in any quantity.

Broad network access. Comprehends the access to capabilities through mechanisms that promote use by heterogeneous client platforms such as mobile phones, tablets, laptops, and workstations.

The established Model also contemplates three categories of Cloud Services:

Software as a Service (SaaS). It represents the capability entitled to the consumer to use the applications provided by and running in a cloud infrastructure. Multiple customer devices and interfaces, such as a web browser or desktop clients, can be used to access the applications. There is no user manager or control of the underlying cloud infrastructure resource (e.g., network, servers, operating systems, storage), but only limited user-specific configuration settings.

Platform as a Service (PaaS). It represents the feature of deploying customer developed or acquired applications in the cloud, relying on provider supported programming languages, libraries, services, and tools. Again, the consumer does not manage the cloud infrastructure, but has control over the deployed applications and hosting environment configuration settings.

Infrastructure as a Service (IaaS). It is the granted customer ability to provision computing resources such as storage, networks, processing, and others, to run arbitrary software, including operating systems and applications. The user does not manage the internal cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control over networking components (e.g., service provider firewall).

Still, NIST [22] listed four Cloud Deployment Models:

Private cloud. In this model the infrastructure is exclusively provisioned for a single organization usage, comprising multiple consumers such as business units. The organization itself, a third party or both might own, manage or operate the infrastructure. It may exist on or off premises.

Community cloud. It is provisioned for the exclusive use of a specific community of users from organizations that have the same demands, such as missions, security requirements, policies and compliance considerations. One or more of the organizations in the community, a third party or some combination of them may own, manage and operate in this model. It may exist on or off premises.

Public cloud. This is provisioned for the open use of the general public. Business, academic, government organizations or any combination of them might own and administrate its infrastructure. It must exist on the premises of the cloud provider.

Hybrid cloud. The infrastructure is a combination of multiple distinct prior cloud infrastructures, private, community, or public. They remain unique entities but are tied together by standardized specifications or technology that allows data or application portability between clouds (for example, dynamic load balancing virtual machine provisioning).

At last, as enumerated by Bohn et al. [23], several roles usually appear for every cloud computing-based system, and the new BaaS solution must contemplate that:

Cloud Consumer. Organizations or people that use the services of a Cloud Provider. For this paper, the Cloud Consumer might use the Cloud to store backups or to make backups from, in a self-service fashion and adherent to every other Cloud requirement.

Cloud Provider. An individual, group or entity responsible for keeping cloud services available to interested parties. In the case of the Cloud Backup architecture, it would host and provide the service and mechanisms where the Could Consumers could deploy, schedule, administer, measure and maintain their backup services.

Cloud Auditor. An organization unity or a third party organization that can conduct an independent audit of cloud services, operations, performance and security of cloud deployment. The backup software must provide the tools that can measure a set of metrics to perform these audits.

Cloud Broker. It comprehends an entity that manages the use, performance, delivery of cloud services, and negotiates the relationship between Cloud Provider and Cloud Consumer. They need a module that can manage the cloud backup service, measure usage, and rate users.

Cloud Carrier. The intermediary connectivity and transport provider from Cloud Providers to Consumers. Usually, they are the Internet providers, and it is the main bottleneck when dealing with inter-Clouds backup. High download and upload throughput are a crucial element for minimum backup compliance.

2.2 Cloud Disaster Recovery Model

Pursuant to Khoshkholghi et al. [12], Disaster Recovery (DR) is a persistent problem in IT platforms, and even more crucial in cloud computing. A Cloud Service Provider (CSP) must find ways to provide services to their customers even if the data center is down (v.g.: due to a disaster). Researchers have shown more interest in Disaster Recovery using cloud computing in the past few years, and a considerable amount of published literature in this area.

Disasters can lead to expensive service disruption [12], regardless of the nature of their causes. A CSP can adopt two different DR models: traditional and cloud-based service models. The use of the traditional model can happen as either a dedicated infrastructure or shared approach. In a dedicated approach, an infrastructure is assigned to one customer, so both cost and speed tend to be higher. On the other hand, in a shared model (also known as a distributed approach) an infrastructure is assigned to multiple users, decreasing both costs and recovery speed.

Figure 2.1 is adapted from IBM [24] White Paper (2012) [1]. It shows that the Cloud-Based DR model is a way of gaining both dedicated and shared model benefits, serving Disaster Recovery with low cost and high speed. Customers choose the appropriate DR model based on speed and cost.

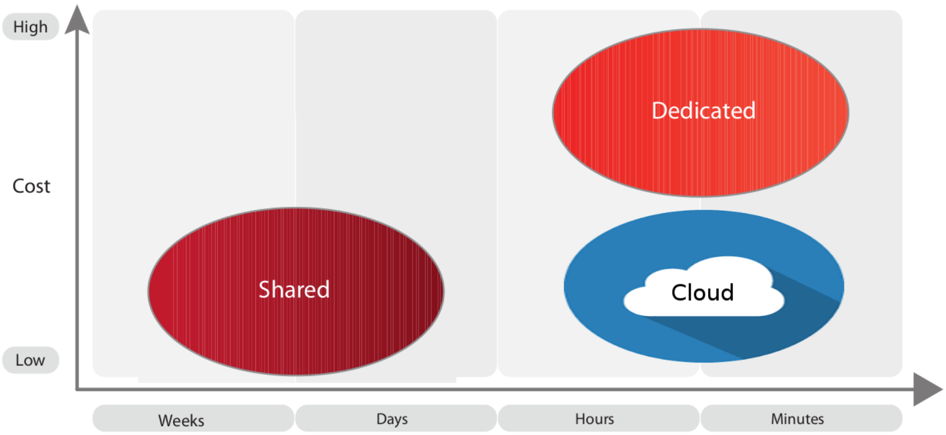


Figure 2.1: Comparison of DR models [1].

Nevertheless, as claimed by IBM [1] the weather causes only 50% of disasters in its cloud, and the rest because of miscellaneous causes (e.g., cut power lines, server hardware failures, and exploitation of security vulnerabilities). In this way, disaster recovery is not only a mechanism to deal with natural events but also for all severe disruptions that may also happen with the modern cloud environment services. It is a critical issue in DR mechanisms is that how can cloud providers tolerate disaster to prevent data loss and service disruption of their data, infrastructure, and services.

Along with Khoshkholghi et al. [12], there are three defined DR levels:

Data Level. The security of application data.

System Level. Reducing system recovery time as short as possible.

Application Level, Application continuity.

2.3 Disaster Recovery Parameters

As stated by Alhazmi and Malaiya [25], Recovery Point Objective (RPO) and Recovery Time Objective (RTO) there are two main parameters that all recovery mechanisms should observe. If RPO and RTO values are lower, the systems can achieve higher business continuity. RPO might be interpreted as the amount of lost data a disaster. RTO consists on the time frame between disruption and restoration of service.

As demonstrated by equation 2.1, the Recovery Point Objective value is inversely proportional to the frequency of backups terminated along the time, where FB represents the *Frequency of Backup*.

$$RPO \quad \propto \quad \frac{1}{FB} \tag{2.1}$$

On the other hand, as exhibited by equation 2.2, Recovery Time Objective formula usually includes a fraction of RPO, the readiness of the backup and five failover steps delays, depending on backup capabilities.

$$RTO = fraction \quad of \quad RPO + jmin + S1 + S2 + S3 + S4 + S5 \tag{2.2}$$

We describe each variable used in the equation 2.2 as the following:

fraction of RPO Computation time lost since the last backup.

jmin Depends on service readiness of the backup.

S1 Hardware setup time.

S2 Operating System initiation time.

S3 Application initiation time.

S4 Data or process state restoration time.

S5 IP address switching time.

Therefore, as alleged by Wood et al. [26], DR mechanisms must have five requirements for an efficient performance:

- Minimum RPO and RTO.
- Minimal impact on the normal system operation.
- Should be geographically separated.
- The application shall be restored to a consistent state.
- DR solution must guarantee integrity, privacy, and confidentiality.

These general requirements are technology independent and can also affect the Cloud-Based Disaster Recovery Models.

2.4 Cloud Based Disaster Recovery Challenges

In keeping with Khoshkholghi et al. [12], a disaster is an unexpected event in the lifetime of a system. It may consist (1) of natural causes, such as tsunami or earthquake; (2) software or hardware failures; (3) human error or sabotage. As claimed by Kashiwazaki [27], it can lead to significant financial or human lives loss. However, only approximately

2% to 4% of the IT infrastructure budget in huge companies are annually destinated to DR solutions [28].

As maintained by Khoshkholghi et al. [12], *cloud-based DR solution is an increasing trend because of its ability to tolerate disasters and to achieve the reliability and availability.* The benefit could be even greater for Small and Medium Enterprises (SME), because of usually shorter resources applied to IT. However, there are notably common challenges of DR in cloud environments:

Dependency. In the opinion of Javaraiah [29], one of the cloud services disadvantages is that customers do not have control of the system and their data, and backup data is on premises of service providers as well. This fact brings customer imprisonment on CSP and concern that higher than expected data loss still might happen.

Cost. As believed by Alhazmi and Malaiya [30], one of the main factors to choose cloud as a DR service is its lower price. Hence, cloud service providers always seek more inexpensive ways to provide DR mechanisms by minimizing the different types of costs that fall into three categories, considered annually:

Initializing cost. Amortized yearly cost.

Ongoing cost. Storage, data transfer, and processing costs.

Potential disaster cost. Cost of recovered and unrecoverable disasters.

Failure Detection. As explained by Khoshkholghi et al. [12], failure detection time strongly affects the overall system downtime. It is imperative to acknowledge the failure as soon as possible to start the DR. However, in multiple redundant sites, it might be difficult to distinguish between a network cloud services access difficulty or a service disruption that would require immediate data restore.

Security. As said by Sabahi [31], well-known TCP/IP protocol used by the Internet and cloud services increases the risks of a malicious user, whether internal or external from organization, obtaining illegal access (e.g.) to a managed virtual machine at the same CSP or physical host machine, in order to exploit vulnerabilities, gain access to other virtual machines and deploy data compromising attacks. Cyberterrorism and Ransomware [32, 33] are growing threats, so protection and recovery of important data will represent an even more important role in Cloud DR plans.

Replication Latency. Accommodated by Ji et al. [34], Disaster Recovery (DR) has two categories of replication techniques: synchronous and asynchronous. Synchronized replication is a high availability mechanism [5] that represent good RTO but bad RPO, since they are usually only able to restore last system working state data, and is not data loss proof (*nothing corrupts faster than a mirror*), is expensive and has higher chance of causing protected system performance impact because of larger overhead. On the contrary, an asynchronous replication backup technique is cheaper, benefits from lower overhead, has higher RTO but lower RPO, because it provides more historical backup versions to restore. Both replication techniques have particular benefits, drawbacks, and purposes. However, this paper focuses specifically on the asynchronous backup software replication requirements.

Limited Data Transmission Capabilities. As pointed by Jian-hua and Nan [35], the remote backup stage is limited by the network connection ability of the cloud. The upload of gigabytes or even terabytes of data between enterprises and cloud storages, or between different clouds, is a great challenge. To reduce the data traffic is also an important means to enhance the service experience, such as backup software both sides deduplication and data compression.

Increasing Data Storage Demand. As specified by Pokharel et al. [36], business snowballing demand for storage is one of the enterprise problems that can be solved by cloud services. It considers the conventional data storage devices replacement with more inexpensive and more flexible cloud storage services. Physical storage, infrastructure management, application interface and access layers constitute the architecture of cloud storage. To satisfy applications and also make sure data is secure, computing might be distributed but storage has to be centralized. In this fashion, the storage single point of failure and data loss are critical CSP challenges that the backup software implementation addresses.

Lack of Redundancy. In line with Khoshkholghi et al. [12], Hua et al. [37], Xu et al. [38], it is a serious threat to the system if it is only possible to store backup data locally, in the same site of primary systems. Original data and backup replicas must be stored in geographically separated places to achieve higher fault tolerance and reliability.

2.5 Cloud Based Backup Architecture

Different proposed solutions are discussed in this section to overcome the cloud-related DR challenges.

2.5.1 Remote Backup to the Cloud

Based on Camacho et al. [39]'s proposal, the need to access backup data any time or from a remote site has grown from a theoretical proposal of a genuine need. They propose a network model where the data is backed up and stored by using an Internet connection on cloud storage.

As demonstrated by Figure 2.2, this is an option for enterprises that have traditional data centers but wants off-load backup work to a cloud.

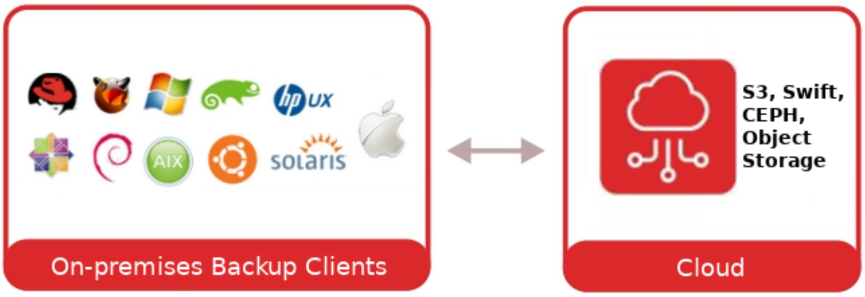


Figure 2.2: Remote Backup to the Cloud Architecture.

Under this architecture BaaS must support Simple Storage Service (S3), Virtual Tape Library (VTL) and other standardized cloud storage technologies. If there are sufficient Internet transmission capabilities, the DR costs should be lower than acquiring local fireproof safes, dedicated link to other sites replication, disk or tape library subsystems.

Similar to this solution, Mao et al. [40] specifies a technique which consists of backup client-side application and the backend service proxy in the customer premises. The later is a layer where the interactions between multiple cloud storage providers happen, hiding the complexity of the backend storage API with the proxy usage.

These solutions attack the following aspects of cloud data protection:

- Cloud dependency.
- Cost.
- Failure Detection.
- Lack of Redundancy.
- Security.

2.5.2 Local Backup from the Cloud

In the light of Ismail et al. [41], Javaraiah [29], this is the simplest of the cloud backup methods, consisting of a solution for cloud dependency problem. The traditional backup software can be deployed on the side of customers to make control get a backup of both data or even complete application using a secured channel (e.g., encrypted). This architecture facilitates migration between cloud service providers or public and private cloud types. In the event of a cloud disaster, the local backup can help to provide the data and systems that were served by the service provider.

As seen in Figure 2.3, this architecture contains a user deployment scenario by incorporating a backup software in the Linux box that will perform the backup of the cloud applications and other data onto local drives.

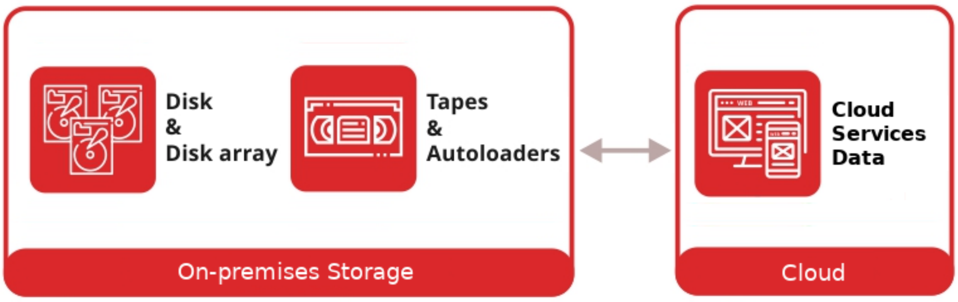


Figure 2.3: Local Backup Architecture.

Still, in consonance with Javaraiah [29], a LTO Tape Drive, a Network Attached Storage (NAS) or another traditional device can be plugged to the Linux box to store backups from the cloud. Also, this solution would attend the different SaaS, PaaS and IaaS recovery needs.

This architecture must be supported by the proposed BaaS, since it address the following important cloud DR challenges:

- Cloud dependency.
- Lack of Redundancy.
- Security.

2.5.3 Cloud Geographical Redundancy and Backup (GRB)

In the words of Pokharel et al. [36], the traditional backup model might use geographical redundancy, but it is usually expensive. It requires additional physical infrastructure, workforce and other resources that make it unaffordable. This architecture proposes a geographical redundancy approach using only cloud systems.

If two cloud zones have a replication of each other, backup of any can be restored if one fails. The cloud multi-site topology, in fact, consists in a ready infrastructure to comply with off-site backups requirement.

As shown in Figure 2.4, there are two zones named Zone "A" and Zone "B". Each zone contains replicas from another. If one zone gets down due to disaster, then a system or data replica can be made available in the other.

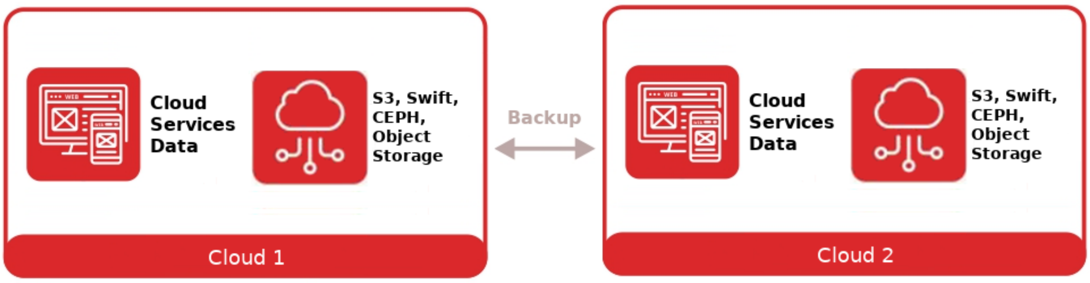

Cloud Services Data
S3, Swift, CEPH, Object Storage
Cloud 1
Backup
Cloud Services Data
S3, Swift, CEPH, Object Storage
Cloud 2

Figure 2.4: GRB Architecture.

An external monitoring unit is required to provide a more reliable disaster recovery system. It monitors the state of both zones, and if it finds there is a disaster at zone "A" or it is in a compromised state, then it alerts zone "B" should restore zone "A" services and data.

Research of Khan and Tahboub [42] has even proposed a method to select optimal locations for multiple back-ups, as an enhancement for GRB technique. Distance and bandwidth are two factors to choose the best sites in this method.

GRB attacks the following cloud backup challenges and therefore should be also supported by the modern backup software:

- Cost.
- Failure prediction and detection.
- Lack of Redundancy.
- Security.

2.5.4 Inter-Private Cloud Storage (IPCS)

Based on Jian-hua and Nan [35], Storage Networking Industry Association (SNIA) recommends at least three backup locations are necessary for business data storage. *Data should be stored in three different geographical locations: Servers, Local Backup Server (LBS) and Remote Backup Server (RBS).* Servers contain original data. Local Backup Server would reduce the RTO and provide independent Cloud communication DR capabilities (private cloud). Moreover, another external private or public cloud consists of the RBS, which protects if a physical disaster (e.g.) affects the original Servers data and LBS.

As shown in Figure 2.5, the IPCS mechanism consists in allowing private clouds to share a single cloud storage service. Multiple private clouds would share the public cloud storage services and it can be called “inter-private cloud storage”. It provides specialized storage services of redundant disaster backup.

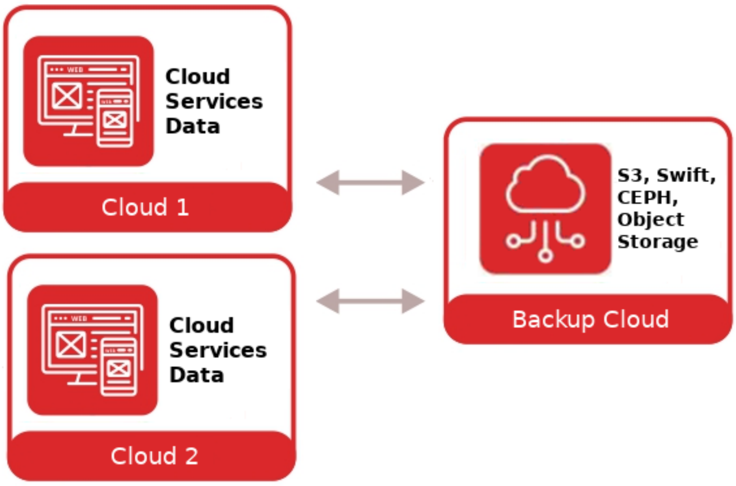


Figure 2.5: The Inter-Private Cloud Storage architecture.

IPCS triangular replication architecture is relatively similar to GBR, except there are three zones to host system or data replicas. In this way, the same cloud backup challenges are addressed:

- Failure prediction and detection.
- Lack of Redundancy.
- Security.

2.5.5 Secure-Distributed Data Backup (SDDB)

As presented by Ueno et al. [2], SDDB data protection technique has six stages:

Initial data encryption. Every data is encrypted when arriving at a data center.

Spatial scrambling. The order of data files is changed by a spatial scrambling algorithm.

Fragmentation and duplication. Data files are divided into fragments, and these are replicated according to the current service level agreement.

Second encryption. Fragments are encrypted again with a different key.

Shuffling & Distribution. Fragments are distributed using a shuffling method to multiple storages with sufficient space.

Transferring Metadata to a backup server. Metadata, encryption keys, shuffling, fragmentation and storage information is saved to a specific supervisory backup server.

In case of disaster, the restore process follows the same inverted steps. The supervisory server will fetch all distributed information, perform the necessary decryption and backup reconstitution, to delivery similar original data to the user.

As displayed in Figure 2.6, the client nodes are composed of PCs, Smart Phones, NAS, and Storage Service. They connect to a Supervisory Server in addition to the Data Center via a secure network. The Supervisory Server (at the right side) acquires the history data composed of the encryption key code sequence (metadata) from the Data Center (at the left side of the picture), using the Internet.

This architecture is more complicated and might not be the desired for all customers, since operations as encryption have a high computational cost, not being suitable for all workloads. Doing it twice might even be considered superfluous from the confidentiality point of view.

In this way, this architecture only addresses the following cloud backup challenges:

- Lack of Redundancy.
- Security.

2.6 Cloud Based Backup Macro Requirements

According to Ismail et al. [41] and under the verified cloud backup architectures, the backup solution must support different backup data, since different layers of operation

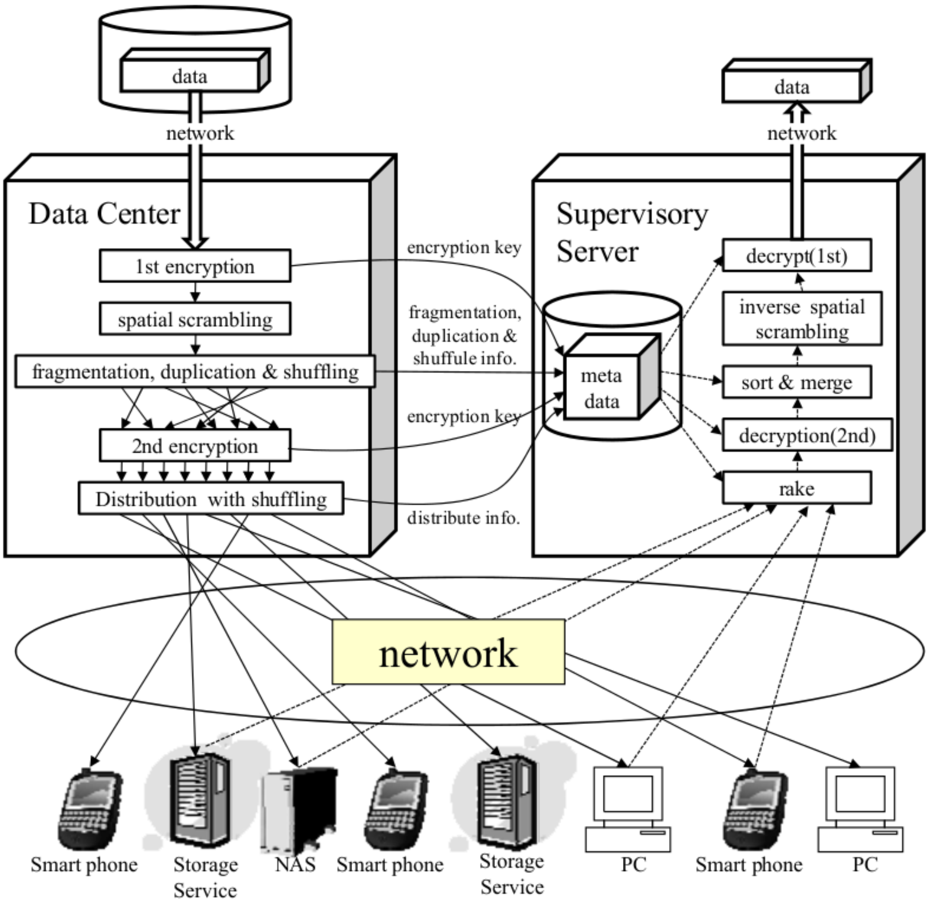


Figure 2.6: Basic configuration of the SDDB Architecture [2].

originate them. Application's specific data such as databases, emails or version control system needs to be backed up within the application itself. File-level - files or folders backup is from file system level. Block Level OS imaging, Continuous Data Protection (CDP) or volume, i.e. Logical Volume Management (LVM).

Still, to offer backup as the XaaS model, it must inherit the cloud characteristics such as self-service, on demand, and ease-of-use. From the technical aspect of the solution itself, the design must be scalable able to add more resources when needed, multi-tenant, able to support multiple users securely (multi-user), and loosely coupled abstraction between service and the infrastructure. The behavior of one part shall not affect the whole system.

2.7 BaaS Provider Requirements

Still in line with Ismail et al. [41], Rimal et al. [43], these are the desired technical requirements for the proposed BaaS system, as follows:

1. **Autonomy.** The system should be designed to adapt dynamically to changes in the environment with less human assistantship possible. User and services behavior of services can be used to improve the quality of services, fault-tolerance, and security, also automatically. New applications objects, such as virtual machines and databases, should have the option of being backed up automatically by the service.

2. **Cloud Scalability.** Still in line with Rimal et al. [43], scalability with large data set operations is a requirement for Cloud Computing, and they can be horizontal or vertical. Horizontal scalability means load balancing and application delivery solutions, such as backup catalog database. Techniques such as Distributed Hash Table (DHT), column-orientation, and horizontal partitioning, might be used if it became a bottleneck. Vertical scalability is related to computing resources used. If BaaS application does not scale well vertically, it is going to increase the cloud provider costs.

   Any application on a database-centric architecture that is limited in scalability if it partitions its database per application instance. It means there is a problem if the backup system needs to scale more than what a single database can provide. It shall support thousands of clients without requiring a separate catalog database. While DHT, column-oriented store, and other approaches address some database issues of scaling write-heavy applications, they do not support complex joins, foreign keys or reporting. They might be part of the Cloud Backup System metadata catalog architecture, and they can play a vital role to reduce writing huge bottlenecks. Anyhow, they should not be considered a replacement for the relational database.

   As explained by Werner [44], the architect must carefully inspect and forecast the system growth where redundancy is required, and how the system should handle heterogeneity. Architects must be aware of which tools they can use under which condition, and what the common pitfalls are. Cloud-based scaling is mostly dependent on the nature of the applications and the expected volume of usage and is not considered a trivial task. A BaaS system shall be able to support thousands of active backup clients and customers simultaneously.

3. **Self-description.** The backup service user interfaces should depict the contained information and functionality in a reusable and context-independent way. When the service contract is updated, the underlying implementation can be changed simultaneously without reconfiguration. Before being deployed, the service may be validated against each potential cloud in runtime. Self-describing services are an advantage because they can notify the application exactly how they should be called and what format of information they will return.

4. **Fault Tolerance.** This enables the backup system to continue operating in case some of its components fail. Fault-tolerance requires fault isolation to the falling components, in general, and availability of reversion mode. Outages occurrence and durations are the main fault-tolerance parameter.

   In the backup subject, jobs shall be able to be automatically rerun and resumed in case of network or another component failure, including off-line backup clients. The backup job logs analysis should provide self-healing and self-diagnosis. Support to subsidiary Machine learning systems like classification or clustering would be helpful not only for detection of failure but also to determine its possible cause.

5. **Interoperability.** In line with [43], the solution shall adopt an agreed-upon framework, ontology, open data format, open protocols or API that enables easy migration and integration of backup data between platforms or different cloud service providers, securely. Backup storage shall support S3 standard for backup storage; backup volume storage and metadata shall be open.

6. **Load Balancing.** As explained by Rimal et al. [43], load balancing is essential part of Cloud Computing and elastic scalability. Software, hardware or virtualized components can provide it, and consists of a technique of self-regulating the workloads between entities of the Cloud service, frequently meaning the usage of multiple servers, hard drives, network or other computing resources.

   Incidentally, the load balancing redundancy is commonly used to provide failover. The service components are monitored continually, the load balancer is informed and no longer sends traffic to it if one becomes non-responsive. This feature is a characteristic of grid-based computing that has been adopted by cloud-based platforms.

   A load balancer is a crucial BaaS requirement to build dynamic architecture. It provides the ways by which backup server instances can be provisioned and deprovisioned. It may happen on CPU, machine, network, application and even on cloud region or datacenter level. The backup server, metadata and storage nodes should provide load balancing.

   Nevertheless, the BaaS solution shall have the capability to manipulate the client backup job storage schedule and forward it to the most favorable targets by using load balancing policies, automatically.

7. **Multi-party.** Should have feature for multi-user interactions and collaboration, such as backup users diary, application elements custom descriptions, comments, and documentation area.

8. **Multi-Tenancy.** It is the usage of shared resources and a single instance of both the object code of an application as well as its database, to support multiple customers simultaneously.

9. **Optimal Provisioned Infrastructure Usage.** The backup solution itself should utilize provisioned infrastructure to provide the service, since every physical equipment, i.e., servers, network switches, storage are treated as a commodity in an IaaS environment.

10. **Quality of Service (QoS).** It provides the guarantee of service performance, availability, security, reliability, and dependability. QoS requirements are associated with service providers and end-users, as established by SLA. QoS subject management systems to monitor resources, storage, network, service migration and fault-tolerance.

    In the case of backups, there must be a caveat, since the way service provider may still not be able to satisfy the performance levels at all the time due to inherent Internet network smaller throughput and higher latency.

11. **Standard Interface.** As stated by Rimal et al. [43], backup services shall be exposed industry standard interfaces, such as web services, Service Oriented Architecture (SOA), Representational State Transfer (REST) [45] or other proprietary service.

12. **Storage and Data Management.** Cloud computing resources shall be elastic to face the changing conditions. For instance, backup storage mechanisms such as Object Storage, VTL or another cloud storage, shall be supported and resources allocated on the fly to handle the increasing demand. Traditional backup retention policies must also be supported to allow automatic backup volumes recycling after expiration.

13. **Workload Management.** Backup should be able to cope with multiple concurrent backup requests, avoiding impacting the performance to any parts of the infrastructure if possible. Phenomenon such as high network usage, high disk read/write operation, or uncontrolled server load should be balanced with transmission cache, bandwidth limitation option, both-sides deduplication, compression, encryption, and other data reducing or security backup system features.

2.8 BaaS User's Requirements

As established by Rimal et al. [43], Ismail et al. [41], there is a list several important functional requirements which are visible to user, and non-functional (or supporting) requirements which are deem important for a BaaS software, as follows:

1. **Adaptability and Learning.** In the light of Cavoukian [46], cloud infrastructure must handle more services, users, resources, and data, tending to have more complex and detailed controls. In the other hand, the service user often must get familiar with the presented application in a short time frame when trying to deal with clouds, and they must be empowered to execute effective controls over information they own.

   As observed by Rimal et al. [43], if users are not fully aware of BaaS usage they might be exposed to different types of security attacks and would not be benefited with these services. Cloud services low pay-per-use usage means less business volume. There are some mechanisms listed below to facilitate users adaptation and to learn the BaaS:

   - Based on observation data recording and users demands, approaches shall be adopted so the users can easily use the BaaS interface.
   - It should be designed to meet the goal of reconfiguration, personalization, and customization, which are Architectural Requirements for Cloud Computing Systems.
   - The backup user interface shall contain an interactive demonstration or illustration that are helpful to unaware users about the application operation. Machine learning techniques can be used to detect patterns of user behavior, chronic operation problems, and difficulties.

2. **Automation.** The backup solution should be easy to use and provide a level of automation which facilitates the execution of the backup process (as "hands-free" or guided as possible).

3. **Backup Standardization.** The BaaS software should provide the same backing up process regardless of OS flavors running within the infrastructure[41]. It should minimally cover most used Unix, Linux and Windows families. As an example, to ensure the consistency of a Windows machine, the file system with Windows Volume Shadow Copy Service (VSS) or have an almost similar method of getting Operating System (OS) consistent state.

4. **Data Consistency.** Backup data should be in a consistent state to perform a successful restore. If any given time backup is performed, the contents of the files

must reflect the content when it happens. It implies any virtual machines transient writes resides in the cache should be flush to the file system before performing backup[41], and the backup software should start other required application-specific preparations. Conforming to VMWare datasheet [47], there are three types of consistencies:

- **Application specific consistency.** Provides application-specific consistency, meaning its data is transformed into a consistent state before the backup execution. Databases require specific techniques[48, 38, 5], such as hot backup mode setting and dump export. In line with Preston [5], operating systems as Windows requires a special *System State* mechanism to allow full OS restore consistency. Each software has one or more methods of producing consistent backups.
- **Crash consistency.** The hypervisor server will flush any transient writes of the guest OS while suspending any further writes for a few seconds, creating a snapshot of the VM disk and aiming disk consistency guarantee, to achieve consistency state.
- **File-system consistency.** The hypervisor server will snapshot the VM file system and save it together with the redo logs which contain latest writes to its virtual disk image files to perform a virtual machine backup. During the restore, the snapshot and the redo log will be committed and written to the disk image file. In line with [49], Physical machines file-systems snapshot is used to provide a similar mechanism. Both crash and file-system consistency address block level only.

5. **Data Integrity.** Keeping up with Ismail et al. [41], data manipulation during the backup process should not tamper data. As an example, a Virtual Machine (VM) disk security copy shall not modify any parts of backup content.

6. **Firewall Protected or NATed Clients Backup.** Traditional backup software initiates the connection vector to the backup server. However, as stated by Stiemerling [50], firewalls and NATs are the most numerous intermediary devices that can impact traffic on the Internet, therefore very hard and laborious to open local premises network ports to every machine that needs backup. Backup clients behind firewalls or with NAT addresses shall be able to transverse these devices, starting the connection and backup jobs to the remote backup service themselves over the Internet.

7. **Non-disruptive Backup.** The solution should always try to reduce or eliminate the backup workload impact caused by operations such as copying, compression, deduplication [41] and restore on the hypervisor, virtual machine or application. If these performance penalties are alleviated, the backup window requirement is considered less relevant.

8. **Service Level Agreements (SLAs).** As established by Rimal et al. [43], SLAs are the contracts that regulate the delivery of services according to predefined parameters. The BaaS software should provide ways of measuring and reporting the backup service delivering, its performance, and to customize elements defined as thresholds to represent different SLAs that the cloud provider might determine.

   Since some performance issues might not be of cloud provider responsibility, Backup-as-a-Service system should also be able to display warnings when applied to these scenarios.

9. **User-Centric Privacy.** Conforming to Cavoukian [46], Cloud Computing end users main consideration regards to the storage of personal or enterprise sensitive data privacy, since data remains at third-party data-centers located around the world. In this environment, privacy on Cloud is a major issue. There are current technologies that can assure data integrity, confidentiality, and security for the clouds to be trusted, and that also applies to backup replicas. These technologies include: encrypting stored backup data, so the cloud provider or third parties cannot read the sensitive data, and virtual LANs or communication encryption to offer safer backup transfers.

   These technologies are quite mature, with some processing overhead for encryption at the client side as a trade-off. Encryption keys storage and some replicas also become critical to be able to restore encrypted backups, but the whole process is transparent to the ordinary user. These should be supported by the BaaS solution as well.

10. **User Consumption-based Billing and Metering.** As explained by Rimal et al. [43], the individual end user consumption-based billing and metering in cloud systems, in an analogy, is similar with the consumption measurement and allocation of water, gas, and electricity consumption costs. These metrics serve as input for the cost management to make planning and controlling decisions feasible and also allowing to check the relation between utilized resources versus costs, break-down analysis, and other techniques. The users need transparency of consumption and billings. BaaS should also report the frequency of services usage to justify the costs.

Conforming to Louth [51], Activity Based Costing (ABC) is a methodology that the user can use to estimate how much their implementation will cost regarding Cloud Computing charges and to test cost transparency.

11. **User Experience (UX).** As reported by Rimal et al. [43], UX is a definite trend of Cloud Computing and comprehends the notion of providing insights into the application end-user's needs and behaviors, to maximize the usability, productivity, and desirability. The user interface shall be simple, uncluttered and designed according to the expected user workflow.

    Usability engineering, Human-Computer Interaction, and ergonomics are considered key requirements to design UX-based cloud applications, and there is no reason to avoid implementing them on the BaaS solution, as the following mechanisms:

    - Analysis of service possibilities, estimation of future usage problems, description of the logic behind the solution by implementing UX design patterns to the backup console interface.
    - End-user involvement in design, conducting early and frequent demonstrations of the user interface.
    - Ensure to consider UX since the beginning of service or product life-cycle.
    - Identification and classification of services from users and the business point of view, according to their potential.

2.9 BaaS Candidate Solutions

Considering the prior studied requirements and as shown in Figure 2.7, there are currently on-premise self-hosted data center backup solutions that can evolve with some development support towards a BaaS delivery solution for CSPs, but we considered other relevant open source based solutions other than the Magic Quadrant.

We include other popular backup software solutions in this survey due to the fact they are trendy worldwide and open source code based. Also, they have open backup storage format and database formats. That are considered powerful benefits since they allow the development of cloud integrations such as custom user reports and data mining.

As displayed in Figure 2.8, the Bacula backup software is the 5th most popular of the verified solutions, making it accredited to be evaluated by the survey. Also, another solution that shares of the same common source code origin (Bareos) is analyzed.

The primary criteria to choose one of the solutions would be the integration with cloud storage; the variety of plug-ins for specific application backup; standard industry

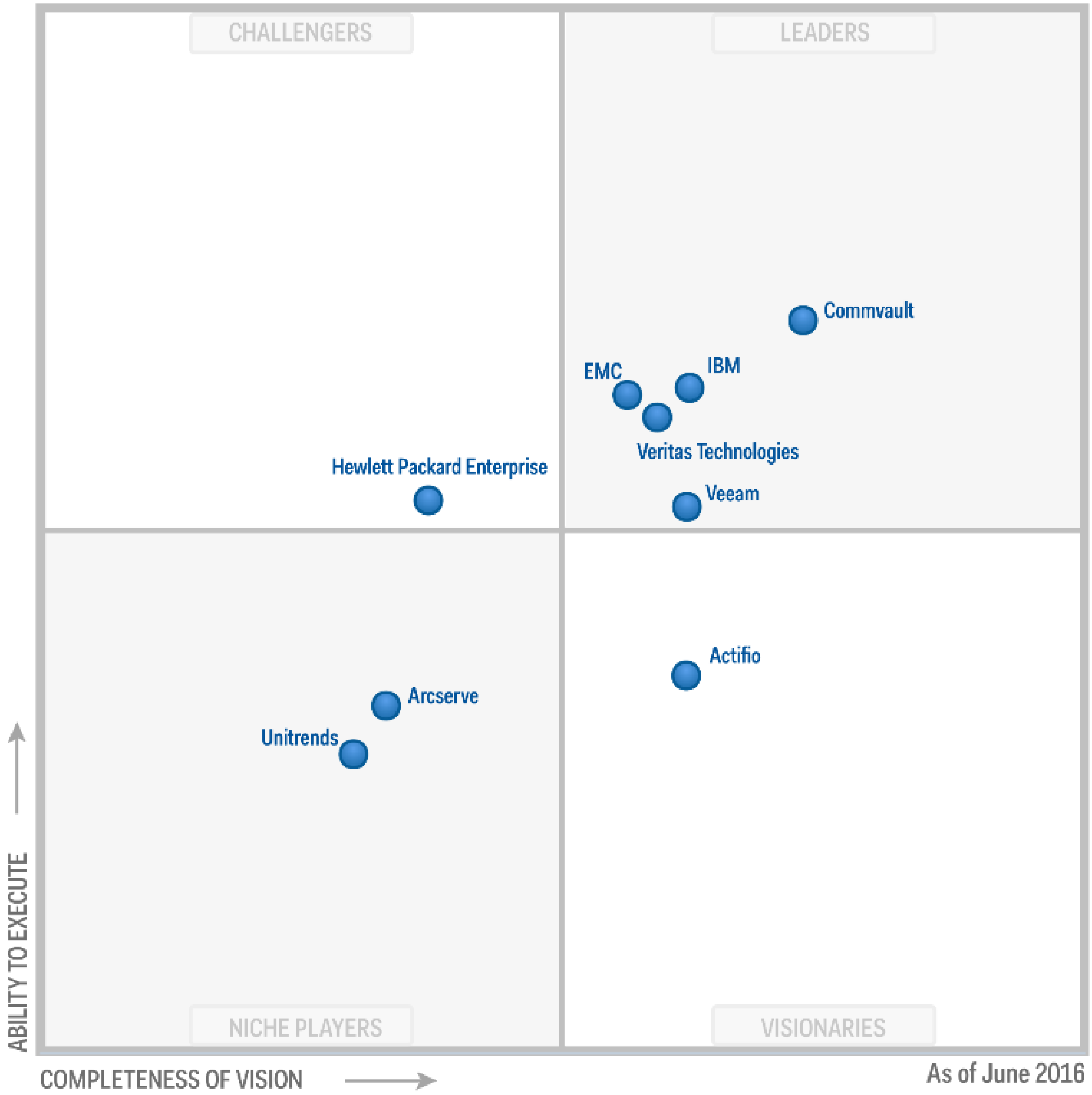


Figure 2.7: Gartner's Magic Quadrant for Data Center Backup and Recovery Software [3].

interfaces such as REST API and open backup catalog format for more natural BaaS development; the licensing cost and the will to collaborate towards a self-service BaaS portal development.

The backup software algorithms such as transmission, compression and deduplication ones exists for a long time and are relatively, and backup performance is largely dependent on hardware infrastructure, storage media and application type backup. Regardless of that fact, we informally gathered a few considerations about the solutions performance and scalability.

Therefore, we state the evaluation of the main considered solutions as follows.

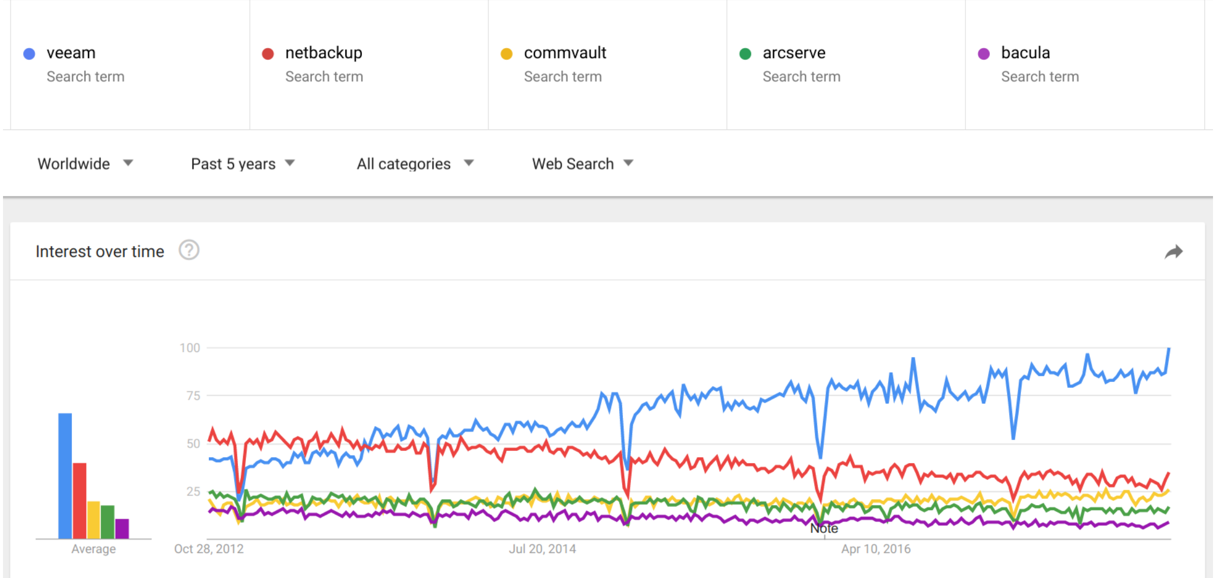


Figure 2.8: Backup Software Solutions Google Queries for the Last 5 years [4].

2.9.1 Arcserve

As explained by Russell et al. [3], Arcserve company focus in its Unified Data Protection (UDP) solution, which is the strategic platform and is offered as software or as an appliance. ARCserve is working to unify backup capabilities and offer a single point of management for UDP.

As listed by Arcserve [52], there are only four supported backup application plugins (Oracle, MSSQL, Exchange and Active Directory), which makes it the worse of the compared solutions on this topic, even though it remains still a popular solution. There is no self-service portal.

As advertised in [52], the socket licensing model results in a price of U$ 7.59 and U$3.08 per client per month, for one and three years respectively, considering a 1/10 socket machine ratio.

According to Russell et al. [3], Arcserve may require further validation for very large enterprise deployments, as the vendor is most typically deployed in the midsize enterprises.

2.9.2 Bacula Enterprise

Bacula Enterprise is a backup software developed by Bacula Systems [53] that derivatives from the Bacula Open Source Community version, the most popular solution of this kind worldwide, according to Google Trends [4]. Despite not being mentioned in the Gartner [54] Magic Quadrant report, unique features such as the REST API cloud industry standard support, open format backup storage and catalog (e.g., allowing backup metadata machine learning [14]), facilitates the development of more advanced cloud features.

Conforming to Bacula Systems [55] Bacula Enterprise has a large number of different applications backup support, including common databases backup plugin such as MSQL [56], MySQL, PostgreSQL and Oracle, and virtual machine hypervisors such as Vmware [57], Xen, RHEV and KVM [58], allowing BaaS to attend a wider range of customers. There are 17 plugins for backing up specific applications.

Bacula also provides a native driver that allows Object backup writing, enabling the backup system to fetch all the theoretical benefits of cloud Storages such as Swift, CEPH, and S3. These characteristics make Bacula Enterprise an excellent candidate to underlie the proposed BaaS solution. However, there is no currently available self-service portal.

As informed by the developer, Bacula Enterprise costs for CSP can go from U$ 16.90 to U$ 3.90 USD per backup client (physical or virtual machine) per month, depending on the number of machines (more clients are more inexpensive), support and all plugins included, which was also considered affordable for this project.

According to noa [53], Bacula Director (server) was tested with more than 10.000, and the storage node can attend to up 600 backup clients.

2.9.3 BareOS

Bareos (Backup Archiving Recovery Open Sourced) is a cross-network backup and recovery data software for most all well-established operating systems. Emerged from the Bacula open source project in 2010, Bareos is developed as a fork and with some new features. Its source code has been published on GitHub [59] with an AGPLv3 license.

According to the developer [60], Bareos only relies on five specific applications backup plugins (LDAP, MSSQL, MySQL, Postgresql and Vmware Vsphere), being considered very limited concerning the number of out-of-the-box integrations. Bareos does not have a self-service portal nor a Global Deduplication feature.

As advertised at Bareos.com [61], license cost per client can range from U$ 52.45 to U$ 5.46, including support, but some plugins are charged apart and can be as expensive as U% 24.00 per application instance. It was not considered a good candidate for this project, both because of technical features and because of pricing.

No scalability numbers were found for BareOS, but since it is a Bacula fork, it should have similar numbers.

2.9.4 Commvault

As displayed in Figure 2.7, Commvault [62] is considered the leader of data center backup solutions, having currently the largest collection of plug-ins for specific applications backup (more than twenty [55] and Cloud Object different storage support. There is also a self-service Portal that allows ordinary users to download backup clients and manage its backup. Only considering technical features, it is probably the best candidate to immediately deploy a BaaS solution nowadays.

According to Commvault [63], however, the license price starts from U$ 31,66 to U$ 12,71 per backed up machine per month, going more inexpensive with more machines hired (simulated up to a maximum of 2500). Higher cost per client is an obstacle for this project.

As claimed by [64], the Commcell backup server can attend up to 20,000 backup clients, and the storage node up to 2,000. These are high values when compared with another solutions.

2.9.5 EMC

In the words of Russell et al. [3], EMC's backup portfolio is a potpourri of separate products acquired over the past decade. EMC flagship solution is the Data Protection Suite (DPS), with Avamar and NetWorker being the two key components focused on the Data Domain Boost for Enterprise Apps. It backs up enterprise applications to EMC Data Domain hardware since a high percentage of EMC backup customers use Data Domain in their environments.

As specified by EMC [65], its product does not support cloud devices as backup storage, which consists of a severe drawback towards a BaaS model. There is no self-service portal, in fact, customers implementing multiple products in a DPS edition need to use multiple user interfaces to perform different operations [3]. Management and support is an issue, with overlapping products and multiple management consoles. Prospects not planning to adopt Data Domain appliances should be aware that the value propositions are weaker without EMC equipment.

There is support for nine different enterprise application backup, being considered moderately limited for full BaaS offer. Monthly clients price begins with U$ 38.25, but every other feature, even storage nodes, and backup server are charged apart. Total cost for a full backup is many times greater.

According to EMC [66], the maximum supported backup size per node is 7,8 TB, and 72 is the number of the maximum number of concurrent jobs [67]. These numbers inform a very limited scalability when compared to the other solutions, only suitable for small to medium business.

2.9.6 Spectrum Protect

As explained by Russell et al. [3], Spectrum is a rebranding of IBM storage solutions. IBM Spectrum Protect and IBM Spectrum Protect Snapshot were formerly known as Tivoli Storage Manager (TSM) and Tivoli Storage FlashCopy Manager, respectively. Tivoli protect focuses on deduplication, scalability and forever incremental backups features.

IBM advertises that, unlike most other backup products that still require a species of periodic full backup even when doing synthetic backups, had a correct incremental-forever methodology since the product's inception. Spectrum protect also supports cloud integration for common S3 storage.

As specified in IBM [68], the IBM Protect has only ten common application plug-ins, with limited support for virtual machines backup (only Vmware and Hyper-v) and databases (only Oracle, Microsoft SQL and DB2). There is no mention of Active Directory or Open LDAP backup support for instance. Also, there is no self-service portal, what keep us from using this solution for the BaaS project.

As verified in IBM [69], IBM protector on the Advanced per socket subscription has a price range of U$ 146.67-90.66 per client per month, considering a ratio of 10 clients per socket. The lower price refers to a five-year licensing package, but still, it is the most expensive of the compared solutions and not eligible for further consideration.

According to Russell et al. [3], a single instance of Spectrum Protect can backup 4 to 5PB of deduplicated data, and Oehme [70] states that thousand of simultaneous connections are supported. It consists on a highly scalable solution.

2.9.7 Veeam

Veeam Backup & Recovery software [71] is another verified solution that is developed by Veeam Software [72] information company. The name "Veeam" comes from the phonetic pronunciation of the letters "VM" as in virtual machine, although it supports backups of other applications such as databases.

Veeam has a limited number of supported application backup: Microsoft Active Directory [73], Exchange [74], SharePoint, MSSQL, Oracle, Vmware and Hyper-v [75], being seven in total. That is a considerable restriction to the number of cloud users that also have other popular applications that must be backed up by the BaaS solution.

As stated by Russell et al. [3], Veeam lacks cloud API integration, therefore discarded as a technically viable solution. Anyhow, Veeam licenses have a flat rate of U$ 12.25 per client, per month, including support and considering an average of 10 machines per socket for comparison basis [76].

According to Veeam [77], the backup server configuration is 1 CPU core (physical or virtual) and 5 GB RAM per 10 concurrently running jobs. In thesis it is possible to run 160 concurrent jobs per server machine with 16 cores, but this would have to be tested since no other research validation information was found. Can be considered scalable solution if numbers are confirmed.

2.9.8 Veritas NetBackup

Veritas NetBackup [78] (called Symantec NetBackup prior to Symantec's divestiture of Veritas) is an enterprise-level backup and recovery suite. Just like the other studied solutions, it provides cross-platform backup functionality to a large variety of Windows, UNIX and Linux operating systems.

NetBackup features a central master server which manages both media servers (storage node) and clients. Central backup server supported platforms include Solaris, HP-UX, AIX, Tru64, Linux, and Windows.

Despite having agents to automate the backup of 14 popular applications, such as databases (Oracle [79], Microsoft SQL and DB2) and Virtual Machines (Hyper-v and VMware), the popular open source databases such as MySQL and PostgreSQL lack of support is a significant downside. As pointed out by Russell et al. [3], there is also no currently CEPH/Swift cloud integration or exploitation, what makes it technically unusable for this project.

As estimated by vendor pricing lists [80], NetBackup licensing price per client is estimated from U$ 332.91 (in the 5 Client starter backup) to U$ 10.58 (standard client), per month and client. These prices are just a conservative estimation, since extra feature modules (e.g., tape library support), storage size and application backup plugins are charged apart and can multiply these numbers. Netbackup verified licensing costs and complicated licensing also makes it unsuitable for further consideration.

As stated by Russell et al. [3], NetBackup is scalable to very large enterprises. As specified by], the solution backup server (master) can support more than 10,000 clients, but no information could be found about the recommended workload per storage node (media server).

2.9.9 Comparison Table

According to the realized survey, Table 2.1 contains the summary of the evaluated characteristics that indicates the currently best backup software for the development and deploy of a BaaS solution, as follows.

Table 2.1: BaaS focused Backup Software Comparison

Solution	Rest API Support	S3 Storage Support	Self-service portal?	Number of Plugins	U$ Monthly Price per Client
Arcserve	No	Yes	No	4	3.08-7.59
Bacula Enterprise	Yes	Yes	No	17	3.90-16.90
Bareos	No	Yes	No	5	5.46-52.45
Commvault	No	Yes	Yes	22	12.71-31.66
EMC	No	No	No	9	38.25 or more
Spectrum Protect	No	Yes	No	10	90.66-146.67
Veeam	No	No	Yes	7	12.25
Veritas Netbackup	No	No	No	14	10.58-332.91

Even though we flagged the Bacula Enterprise as not having a self-service portal, its manufacturer (Bacula System) proposed to support the development a cloud multi-tenant self-service portal during this research according to the requirements that are fixed by this study, what happens to be a favorable element.

The Bacula Enterprise Cloud S3 Object backup storage support, REST API availability, open catalog format, one of the broadest range of applications backup plug-ins and the reasonable pricing, makes it the most suitable backup software to deploy a minimum BaaS product in a short time frame, with the bonus of having open format backup data and metadata.

Nevertheless, we plan to integrate the BaaS developed portal using REST API, so it is also likely to work with the Bacula open source version and even other backup software that might happen to have that cloud industry feature, even if some command changes are needed. In our opinion, that adds value to the proposal.

2.10 Chapter Summary

This chapter verified the new cloud reality disaster recovery and backups problem. We described the parameters, challenges and most prominent architectures. We have surveyed the Cloud Provider BaaS's Systems and Users Macro Requirements. There is another survey about the current candidate BaaS software solutions, a list of desired criteria and a comparison between them.

The next chapter details the chosen Bacula Enterprise solution required development effort to achieve a BaaS solution and the design of the prototype that will be used to attend the "Remote Backup to the Cloud" architecture.

Chapter 3

The Bacula

In line with Chapter 2, the studied cloud backup challenges, architecture, and requirements will now subsidize the proposal of a prototype development that will serve to one of the researched cloud backup structures.

Bacula [81] is a network distributed open source backup software [82] that attends part of the established architectures, users, and technical requirements of this study. We selected its more advanced version, Bacula Enterprise [83], as the overlying backup system for the development of a BaaS interface.

Conforming to Google Trends [84], Bacula is the fifth most popular enterprise multi-platform server backup system worldwide. It supports Full, differential, incremental, copy and migrate multiplexed jobs [85]. Backups are written into tapes or disk files, using open format bytes-sequence volumes. Compression, block deduplication, backups from FIFO, communication and data encryption are other currently deployed features.

As shown in Figure 3.1, Bacula follows the classic backup software modules distribution. There is a central backup server called *Director*, specific device storage service nodes named *Storage Daemons* and backup clients for different operating systems are known as the *File Daemons*. A database named *Catalog* stores backup metadata, such as job logs, termination status, list of copied files with paths, storage media association, and other information.

As displayed in Figure 3.2, the backup system administrator sends user commands to the Director daemon, such as start a backup job (or schedules it previously). The Director sends software commands to the desired File Daemon (backup client). The File Daemon copies, compress and encrypt (if enabled) files and data to the selected Storage Daemon. Storage Daemon writes backups to disks, magnetic tapes or other devices, and send files metadata to the Director that also stores them in the Catalog database.

According to Sibbald [86], Bacula was the first published backup software to use a Structured Query Language and supports MySQL [87] and PostgreSQL [88] different

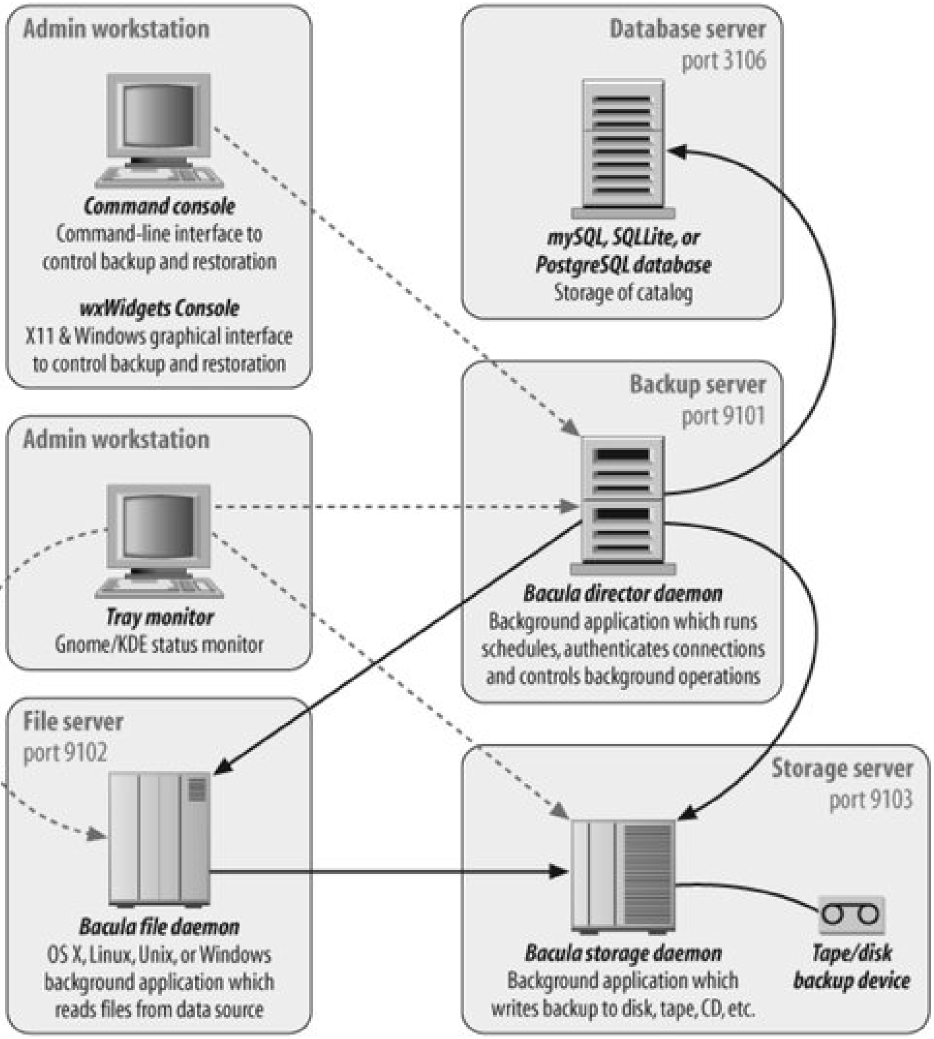

Admin workstation
Command console
Command-line interface to control backup and restoration
wxWidgets Console
X11 & Windows graphical interface to control backup and restoration
Database server
port 3106
mySQL, SQLLite, or PostgreSQL database
Storage of catalog
Admin workstation
Tray monitor
Gnome/KDE status monitor
Backup server
port 9101
Bacula director daemon
Background application which runs schedules, authenticates connections and controls background operations
File server
port 9102
Bacula file daemon
OS X, Linux, Unix, or Windows background application which reads files from data source
Storage server
port 9103
Bacula storage daemon
Background application which writes backup to disk, tape, CD, etc.
Tape/disk backup device

Figure 3.1: Bacula Services Distribution and Network Ports [5].

open database Catalog service.

Consistent with Zhang et al. [89], Bacula copies backup data from servers and stores it in the Storage Daemon, with no index. Even for disks, read and write operations are made in similarity with a tape library, when each recovery or backup is sequential. There is no distinction between a disk or tape from the storage device point of view so that the same volume can store data from different platforms. Therefore, backed up files metadata is also stored in the database for quick restore job files selection.

As understood from Sibbald [90], Bacula has two backup data writing format: traditional non-deduplicable, and a newer deduplicable one. As seen in Figure 3.3, by default, Bacula Storage Daemons writes different servers backup metadata, headers, and file data to a single file compacting and mixing them. It makes the default Bacula volumes unique and challenging to deduplicate.

In this way, Sibbald [90] proposed and deployed a new optional Bacula deduplicable

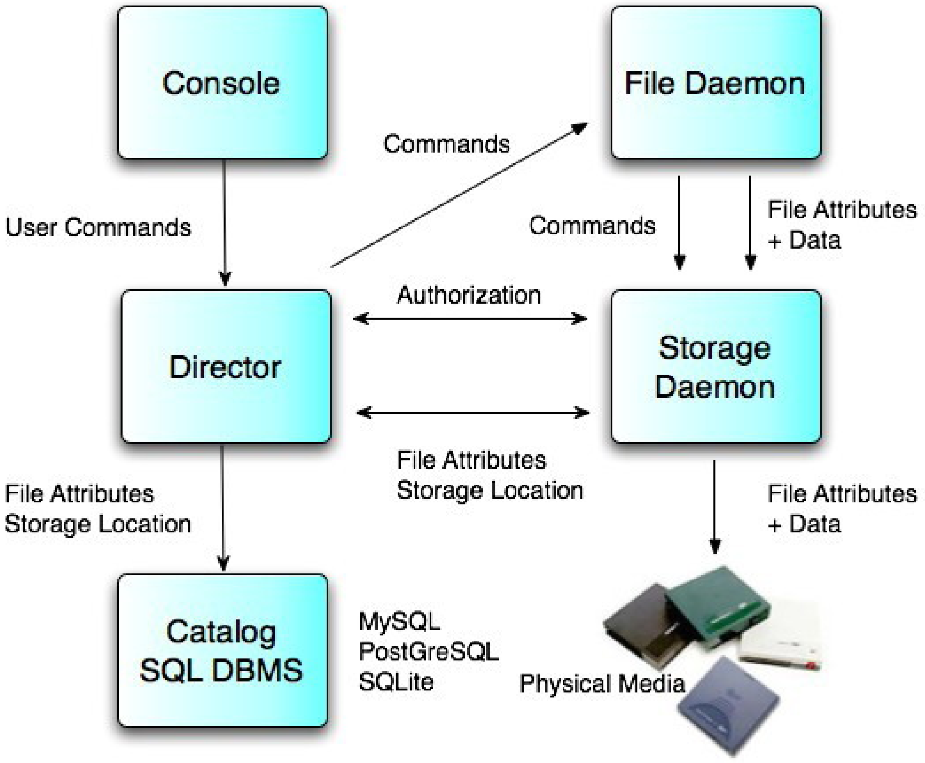


Figure 3.2: Bacula Traffic Characterization [6].

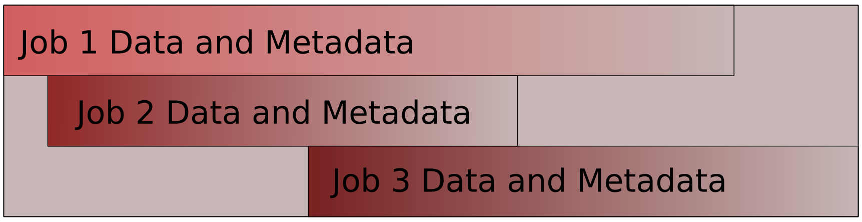


Figure 3.3: Bacula traditional volume writing format.

archive volume as drawn in Figure 3.4. It is a data container from an original stream of multiple files, and which can be optimally deduplicated by an underlying deduplication storage system. The method comprises receiving data records representing metadata and backup data, at least a part of which is already separated. The the first file called metadata volume contains backup metadata, header data and references to the paired data volume. The second file called Aligned Volume only contains backup data, increasing block-level deduplicability.

As reported by Xia et al. [91], the duplicable Bacula format allows reducing backup storage or archiving datasets size by a factor of 4-40X using data deduplication. This data reduction technique is vital to save disk and cloud object storage resources. For magnetic tapes, sequential access makes virtually impossible to deduplicate written data.

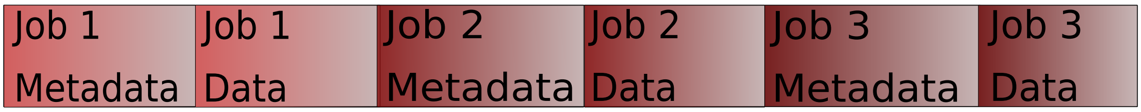


Figure 3.4: Bacula Aligned Volume deduplicable writing format.

In keeping with Sibbald [92], however, Bacula also provides its Global Deduplication mechanism, that is capable of performing duplicate check in backup datasets from all clients before data transfer. That is very useful for clients in the cloud, for example, reducing WAN data transfer.

3.1 Bacula Features

As enumerated by Bacula Systems [93], Sibbald [92], ahead are the main Bacula backup current features. We divided them into categories for better organization, as follows: General, Job, Interface, Catalog, Storage, Data Reducing and Specific Application Backup. We also associated each feature with the studied Cloud services macro-requirements.

3.1.1 General Bacula Features

General backup system features are usually common to all enterprise backup software and are related to the implementation of basic concepts such as job, volume, pools, policies and holistic behavior, such as distributed components structure and scalability.

- Volumes can contain jobs from different dates on the same or in multi-volume saves. Bacula automatically requests the next Volume when the first gets full, and continues the backup Pool and Volume library management, providing Volume flexibility - e.g., monthly, weekly, daily Volume sets, Volume sets segregated by Client (automation).

- Automatic backup volume retention times management and recycling (automation, storage, and data management).

- Bacula labeled volumes, preventing accidental overwriting. Support to IBM/ANSI tape labels, which are recognized by many enterprise tape-managing software (backup standardization and interoperability).

- Platform independent volume data format. Linux, Solaris, and Windows clients can all be backed up to the same Volume if desired. The same for restore operations (backup standardization and interoperability).

- Open format backup volume storage (backup standardization and interoperability).
- Backup jobs load Balancing between storage devices (automation and load balancing)
- Distributed Ethernet network daemons, with a centralized Director - backup server (optimal provisioned infrastructure usage).
- Multi-threaded implementation (optimal provisioned infrastructure usage).
- Efficient use of provisioned resources such as network, CPU, and storage (optimal provisioned infrastructure usage).
- Client initiated backup option, even behind firewall or NAT (optimal provisioned infrastructure usage).
- Any number of Jobs and Clients can be backed up to a single Volume (optimal provisioned infrastructure usage).
- A flexible message handler includes message routing from any daemon back to the Director and automatic email reporting for backup system users (automation).
- The volume data format is upwards compatible so that Bacula newer releases can read old volumes for 30 years (interoperability).
- Bare Metal Recovery plugin for Linux and Windows systems, with UEFI and EFI support (interoperability).
- Bandwidth limitation option (workload management).
- Snapshot technology and Snapshot Management to backup or get a read-only copy of a File System data set, allowing running applications to continue writing their data (workload management).
- Support to tiered backups such as disk-to-disk-to-tape technique (workload management).

3.1.2 Backup Job Features

Backup jobs are the backup system operating unity and they always have finite duration [13]. They can write a list of files or folders from the backup clients to the storage node (backup), or they can retrieve data from the storage to the backup clients (restore). There are also other job types, such as copy and migration (used to make a second copy or move already terminated backup jobs for another storage), and verification jobs (used to verify consistency from different backup elements: storage, catalog, and original client data).

- Internal scheduler for automatic Job execution (automation).
- Multiple independent multi-platform job configuration (automation, interoperability).
- Backup and restore clients of any type ensuring that all specific system attributes of files are properly saved and restored (backup standardization).
- Job sequencing using priorities (workload management).
- Incomplete jobs resume. Stop (pause) and restart jobs commands.
- Concurrent multiple Jobs execution option (multiplexing).
- Differential, Full, Incremental, virtual-full (synthetic full) files backup and recovery functionality (workload management).
- Progressive Virtual Full - only changing Catalog indexes workload management).
- Copy and Migration support: move backup data from one pool or volume to another (workload management).
- Always back up a specific file optional setting (workload management).
- Backup jobs timeout, retry times and interval configuration; local cache for remote object storage writing (fault-tolerance, workload management).

3.1.3 Bacula Interface Features

Traditional backup administration interfaces provide ways of visualizing, starting or configure new backup jobs. They consist in GUIs and CLIs, both with desktop and web access options [7]. Few backup systems also support standardized industry interfaces such as Representational state transfer (REST). Bacula supports the following interface related features:

- CLI and different web and desktop GUI alternatives for backup system operation (interoperability).
- REST API to define, easily and with a standard programming mechanism (interoperability, self-description and standard interface).
- Custom Bacula elements descriptions by the user (multi-party).
- Monitoring of job throughput, network interfaces negotiated speed, backup duration time frames, daemons memory usage and software compression efficiency, including estimation to next jobs (QoS).

3.1.4 Metadata Catalog Features

As explained by Guise [7], a backup catalog database is fundamental and critical, necessary to achieve faster recoveries. It should track the backup media, the backup list of files, the stored backups in each volume, which medias are necessary for a given restore, volumes usage information, volume pools, retention and backup policy configuration. The indexes also allow fast-forwarding through media.

We list Bacula Catalog related features as follows:

- SQL standardized metadata Catalog, containing information backup volumes, pools, jobs, and copied files (backup standardization).
- The Bacula Catalog database permits immediate viewing, navigation, and selection of files saved on any particular volume, besides reporting capabilities (automation).
- Allows the restore of single or multiple files selected interactively either for the current backup or a backup before a specified time and date (automation).
- Automatic pruning of the database (removal of old records), simplifying database administration (automation).
- Support for MySQL and PostgreSQL databases, in an open format and extensible user queries (interoperability).

3.1.5 Backup Storage Features

According to Guise [7], backup systems have a master service controller, backup clients, and slave storage nodes. The last ones are responsible for storage information on physical devices such as tape libraries and disk storage arrays. They can also be called media servers or storage daemon, depending on the solution. In the Bacula case, the storage daemon has the following features:

- Cloud storage (S3, CEPH, Swift) backup storage support with local cache (interoperability, fault-tolerance and interoperability).
- Storage Daemon to Storage Daemon (SD2SD) for replication capabilities, with deduplication.
- Storage Daemon calls Backup Client if behind a Firewall (firewall protected or NATed clients backup and interoperability).
- Storage Daemon reporting allows retrieving a range of information such as available disk space and disk usage (data and storage management).

- Storage Daemon also supported on Windows systems (interoperability).
- Support for Windows mount point snapshots (data consistency, data integrity, interoperability).
- Advanced support for most Storage Devices, with SAN shared storage capabilities and NDMP backup (data and storage management and interoperability).
- Tape libraries (media autochangers) support using a simple shell interface that can handle virtually any autoloader program (interoperability).
- Support and automatic tape labeling for autochanger barcodes (interoperability).
- Automatic support for multiple autochanger magazines either using barcodes or by reading the tapes inventory. Support for multiple drive autochangers (storage and data management).
- Data spooling to disk during backup, with the subsequent write to tape from the spooled disk files. This feature prevents tape "shoe shine" during incremental or differential backups (workload management).

3.1.6 Data Reducing Features

In line with Gartner Consultancy Report [94], digital data has snowballed to a level that frequently leads to backup storage capacity depletion. It is imperative to deploy mechanisms such as compression and deduplication to reduce backup software storage demand, without sacrificing data retention or downgrading policies. We list Bacula's data reducing perks as the following:

- LZO [95] light-weight faster compression algorithm for clients with CPU overhead (non-disruptive backup, optimal provisioned infrastructure usage and workload management).
- Bacula software compression support with different GZIP [96] algorithm levels (workload management).
- Communication line only compression allowing, reducing data transmitted across backup daemons when software data compression is not necessary or desired (non-disruptive backup and workload management).
- Global Endpoint deduplication, deduplicating data from the clients - source (workload management).
- Support to filesystem and hardware deduplication - Aligned Volume Format (workload management).

3.1.7 Security Features

Consistent to Ueno et al. [2], a backup system (as any distributed solution) must guarantee, as far as possible, the security of users' or institutes' massive files of essential data from any risks, such as an unexpected natural disaster, a cyber-terrorism attack, interception, hardware failure and others. To meet this need, techniques such as hash checksum verification and encryption must be supported with affordable computing cost to provide high security. Some of Bacula's security features are listed as follows:

- All Volume blocks contains data checksum - approximately 64K bytes (data consistency, workload management).
- Emergency restore application, permitting extraction of files when Bacula system and/or Catalog are not healthy or available (data consistency).
- Contingency ability of recreating the catalog database by scanning backup Volumes (data consistency).
- Supports the following data encryption cipher: AES 128, AES192, AES256 or blowfish, and their digest algorithm (data consistency, user-centric privacy and workload management).
- Support to Windows EFS (data consistency and interoperability).
- Verify the reliability of existing backed up data and backup data corruption. Verification of files previously catalogued backup files (data consistency and QoS).
- CRAM-MD5 password authentication between each Daemon (user-centric privacy).
- Configurable TLS (SSL) communications encryption between each Daemon (user-centric privacy and workload management).
- Configurable Client source data encryption (workload management).
- Computation of MD5 or SHA1 signatures of the backed up files (data consistency and data integrity).

3.1.8 Specific Application Backup Features

In agreement with Guise [7], backups encompass far more than just filesystem-level protection. For many organizations, the most critical information consist of application data and databases. The backup system must be designed to protect these as any of the involved filesystems. We registered the following Bacula related features:

- New applications objects are automatically added to the backup if no filter is configured (autonomy and automation).
- VMware Vsphere client-less Virtual Machine backup support plugin featuring Single File Restore (automation, data consistency and data integrity).
- KVM, RHEV and Xen Virtual Machines backup and restore (automation, data consistency and data integrity).
- Hyper-v Virtual Machines backup and restore Plugin (automation and data consistency).
- OpenLDAP [97] and Active Directory Directory Server plugin (automation, data consistency and data integrity).
- Plugins currently exist for MySQL, PostgreSQL, MSSQL and Oracle databases (automation, data consistency and data integrity).
- Microsoft VSS plugin with support for Microsoft Exchange (automation, data consistency and data integrity).
- Gmail and Zimbra backup plugin
- Custom generic FIFO/Named Pipe backup plugin

3.2 Development Effort to a BaaS Solution

The Table 3.1 lists and analyzes what Bacula software requires to be developed to be considered a BaaS solution that would attend most of the studied Cloud backup architectures. It lists the gap between the Cloud Application requirements and what features representative backup software deliver today. They are necessary to attend the different studied Cloud backup architectures.

3.3 Chapter Summary

This Chapter detailed the Bacula backup system characteristics, features, and the full roadmap to a BaaS integrated solution. Next Chapter presents the BaaS proposal, objectives, development scope for this work, interface topology, main screens, abilities, case study, user roles, the usability evaluation, user-inquiry details and results.

Table 3.1: Roadmap to the BaaS Solution

Id	Requirement	Macro Requirement
1	Shall permit reconfiguration personalization and customization of the BaaS interface.	adaptability and learning
2	Shall contain interactive illustration or demonstration to instruct unaware users.	adaptability and learning
3	Should enable tracking of user interface demands, for observation and machine learning	adaptability and learning, user experience
4	Shall support thousands of end-users and backup clients automatically, with horizontal and vertical techniques.	cloud scalability
5	Should provide backup logs machine learning feature.	fault-tolerance
6	Shall perform self-diagnosis and self-healing.	fault-tolerance
7	Shall have the ability to select one or more multiple most favorable backup storage targets according to policies, automatically.	load balance
8	Shall have a web interface users backup diary or customs documentation module.	multi-party
9	Shall provide the usage of a Bacula single instance of the object code and database to support multiple customers.	multi-tenancy
10	Shall provide monitoring of backup clients average CPU usage.	QoS
11	Shall allow online storage allocation to handle increasing demand.	storage and data management
12	Shall be able to create new tenants easily	automation
13	Shall provide measuring and reporting about the backup service delivering and its performance	SLAs
14	Shall provide ways of establishing thresholds to measured backup performance	SLAs
15	Shall have a resource consumption-based Billing and Metering module, based on Activity-Based Costing	user consumption-based billing and metering
16	Shall have a user-centered design	user experience
17	Shall consider identification and classification of services from the user and business point of view, according to their potential	user experience

Chapter 4

BaaS Solution Proposal for a Local Backup to the Cloud Architecture

Due to the limitation of time, this work must focus on one of the reviewed Cloud backup architectures. According to Heslin [98], self-owned data centers and collocation are still the most common IT infrastructure model used by the companies. Therefore, the proposed solution aims to prototype a BaaS solution for the *Remote Backup to the Cloud*, since it would serve to the majority of enterprises. It is important to highlight this structure also addresses lots of reviewed backup challenges, consequently one of the most relevant.

In this scenario and considering Cloud is user-centric, the most notable and critical requirements are interface related, such as a multi-tenant user-friendly operation. Now the backup Clients are deployed and configured by the ordinary Cloud User, instead of the old specialized backup administrator. The traditional backup software is not multi-tenant, and as observed by Amvrosiadis and Bhadkamkar [99] they are too complicated even for specialized users since the majority of them uses stock backup software configurations for schedules and backup retentions.

That said, we chose to deploy the requirements identified as the number: 1, 9, 12 and 16 from the Table 3.1 in the prototype proposed solution. They should lead to a self-service multi-tenant user portal Cloud Providers can make available to their customers, so they can quickly set up a *Remote Backup to the Cloud*.

4.1 Main Objective

This work aims to deploy a Backup as a Service solution under the Remote Backup to the Cloud architecture. To achieve what we aimed the following specific objectives.

1. To enroll cloud backup challenges.

2. To find currently proposed cloud backup architectures.
3. To determine what BaaS features must be added to the traditional backup software.
4. To deploy, evaluate, and validate the proposed prototype.

We detail the enumerated Objectives in the next subsections.

4.1.1 Objective 1 - To Enroll Cloud Backup Challenges

In the words of Russell et al. [3], Russel [9], enterprise backup is among the oldest, most performed tasks by infrastructure and operations professionals. It is necessary to verify how if and how it remains relevant under the new Cloud model. A survey about Cloud backup challenges shall indicate the backup importance and the main problems that BaaS solutions should address with more or less priority.

4.1.2 Objective 2 - To Find Current Proposed Cloud Backup Architectures

Another survey about currently proposed Cloud backup architectures should provide consolidated directions of how other researchers are addressing the Cloud backup challenges. It should indicate what techniques are being used and will provide scenarios where the new proposed BaaS solution will have applicability.

4.1.3 Objective 3 - To Determine what BaaS at must be Added to Traditional Backup Software

There is currently well established traditional backup software, even considering they have been designed and optimized for outdated environments and use cases [3, 9]. In this way, a review should indicate what features are desired for a BaaS software solution, and which of them still must be improved or developed.

4.1.4 Objective 4 - To Deploy, Evaluate and Validate the Proposal Prototype

The studied Cloud backup challenges, architectures, and software requirements should lead to the planning and deploy of a BaaS solution prototype to attend one of the researched architectures. The experiment will be specified, its performance will be measured and analyzed under a significant and representative workload. Finally, we will present the results and analysis.

4.2 Development Scope

After summarizing the BaaS solution requirements, this section is intended to define the scope of work of the specific application to develop.

The purpose of the development is to provide cloud users a Web Graphical User Interface (bcloud) that aims to be clean, simple and easy to use. Mainly, bcloud must allow the user to download, install and manage Bacula backup clients, perform backup and restore operations, according to the following listed features:

- Shall grant regular username/password access to ordinary users.
- Shall grant user ability to download software backup clients.
- Shall work for both Windows, MacOS, Unix and Linux Clients.
- Shall grant ordinary user ability to request attachment of clients to a specific backup server.
- Shall grant ordinary user ability to create, update and exclude its backup.
- Shall allow ordinary users to restore Jobs.
- Shall allow users to select files to restore.
- Shall allow the ordinary user to start, stop, restart, or cancel backup and restore jobs.
- Shall allow ordinary users to choose alternative authorized restore clients.
- Shall have Clients connection encrypted.
- Shall have users data quota control feature.
- Shall have the option of forever incremental backups.
- Shall have ordinary user-friendly documentation, including the ports used by Bacula for firewall exception rule addition.
- Shall show a dashboard displaying: the date/hour of the last backup, the size of the last backup, all the backup volume to time, last time online, next backup and schedule list (and a link to the scheduling module).
- Should grant ordinary user setting alerts for successful or failed jobs, including the ones after a given time without terminated backups.
- Should show the backup and restore progress meter.
- Should have user quotas management, billing feature and report.

4.2.1 Topology

As shown in Figure 4.1, bcloud topology will rely on the Bacula REST API that enables control of Bacula with high-level HTTPS/REST calls.

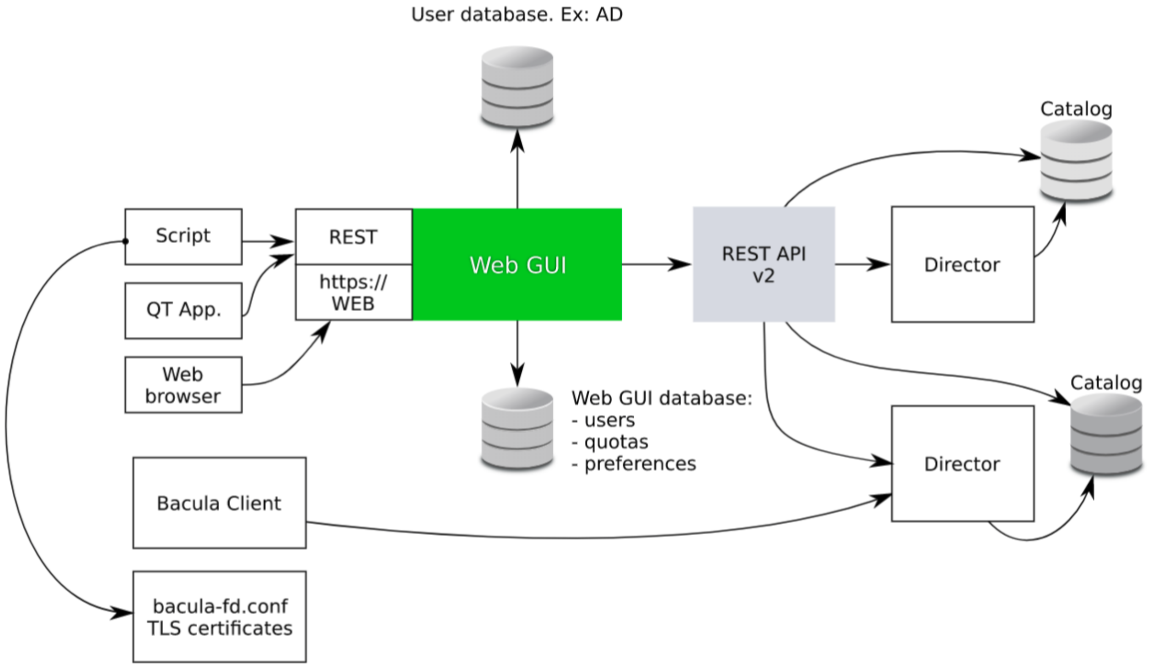


Figure 4.1: bcloud topology

The REST API plays a crucial role on the BaaS interface development, creating an abstraction layer for Bacula operation, granting a faster programming and making unnecessary to know or to change Bacula source code. It would be probably impossible to develop the bcloud interface in such a short time frame without using this industry standard perk.

The API also makes the backup system operation safer, reliable and more error-proof. It allows to view, create, modify and delete Bacula objects, that can be either configuration information in the various component configuration files, or Bacula catalog database information. The API also permits the programmer to use a command interface.

We selected the Smarty PHP framework for development. The use of a popular PHP framework will allow the screens to be customized, and will also allow future addition of custom modules. Some items such as the primary logo, the header, and the footer will be configurable by the customer. The Web GUI will offer the possibility to manage translations (with po files, e.g.). It will also be able to manage multiple Directors (backup servers).

The Tray Monitor is a multi-platform QT program that enables to monitor the Client (File Daemon) and start backup jobs, especially when behind inaccessible firewalls or

NATed networks. The end user will be able to use the Tray Monitor program to start client initiated backup jobs or to perform other Bacula related actions.

4.2.2 Backup Clients Download Center

The bcloud will offer a PHP module (class and templates) that will list a local server directory and the sub-directories for self-service backup clients download. More advanced features such as an upload center, tags, groups and versioning are not part of the first implementation. To make binaries available to the customer, the system administrator will put files in a local directory with a specific path layout.

4.2.3 Automatic Client Registration and Encryption

The bcloud will include tools to register a new Bacula File Daemon (Client). The registration program will connect to the Web GUI via a REST protocol. The registration program will take the following arguments:

- the address to connect to
- the username
- the password
- the client name

At the connection time, the registration program will check the TLS thumbprint of the Web GUI. Once verified by the user, the program will authenticate with the WebGUI, generate the necessary TLS configuration from the root certificate and create Bacula resources and objects needed to register the new client.

The interface will save all the Director configuration in a temporary area (named work set). The Client will be configured with the Director information (Name, Password and TLS certificate). The Client's local daemon will be started. The registration program will have to be executed on the Client with the necessary permissions to edit the daemon configuration file, the Tray Monitor configuration file and control the backup service (root on Unix and Administrator on Windows).

By default, the CSP administrator will be notified about the registration request and can commit, modify or discard the request. A graphical GUI interface allows running the registration program as a wizard. The account used to sign in will be associated with a Bacula Restricted Console and the backup server resources. The name of the resources is explicitly linked with the account name. The administrator can limit the number of Clients created by a user.

4.2.4 Network Configuration

All Bacula Clients that are attached to a Director will be located on networks that may be protected by firewalls or behind routers. A direct connection from the Director to the File Daemon using the TCP port 9102 may not be possible without a VPN configuration.

As presented in Figure 4.2, the client initiated backup and restore features allow jobs to be started from the Client via script or the Tray Monitor program, being able to transverse Firewall and NAT equipment.

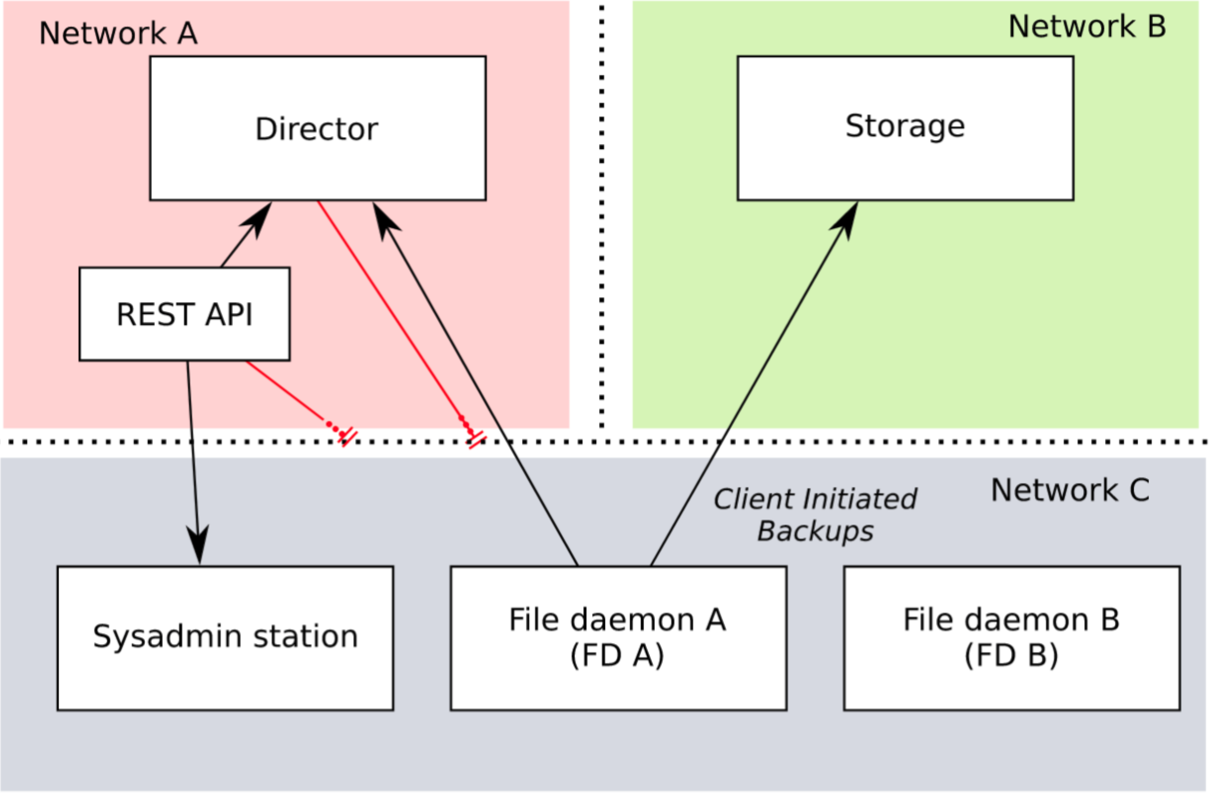


Figure 4.2: Client Initiated Backup

4.2.5 Quota Management

Bacula can limit the number of bytes written by a given Client using the Pool resources. If a Pool can contain 100 volumes, and each volume can have a maximum of 1GB of data, then the Pool will not store more than 100GB of data. Multiple jobs can share volumes for the same customer. When a Pool is full, Bacula can purge and recycle the oldest volume to continue the current job or cancel the current job after a given time.

The bcloud will query the Catalog and send notifications to end users if their Pool reaches a given size. This notification can be done at the end of a Job, or can be scheduled a couple of times per day.

4.2.6 Job Configuration

The end user can submit requests to create jobs for a Client, configure the FileSet and schedule them. For example, “Folder 1” will be backed up at 11:00 every day, while “Folder 2” will be backed up at 13:00 every Sunday. Custom directives will be available via template. The naming scheme used in all user-defined Bacula resources will be predictably fixed.

4.2.7 Job And Restore Management

The end user can monitor and stop a backup job from the bcloud interface. Also, can browse files, select files and directories, and restore to any Client that is accessible.

4.2.8 Billing Feature and Report

The “User Overview” screen includes statistics about the jobs stored in the catalog. Information such as:

- Total number of files
- Total number of bytes
- Total number of jobs
- Total number of clients
- Quota
- Percent of Quota used

More advanced reports can be provided and easily deployed with the REST API backup Catalog query.

4.2.9 List of Screens

As displayed in Figure 4.3, the main page “My Apps” can list multiple kinds of internal bcloud modules or applications. “Download Center” and “My Data” are two examples. We might add other applications in the future.

Elements such as the header, the footer, the colors and the logos will be configurable by the customer. The following screens are planned for bcloud:

Login Screen The login screen will offer standard options. Input fields will depend on the authentication module. The Password Reminder feature is not available with some authentication backends such as Active Directory.

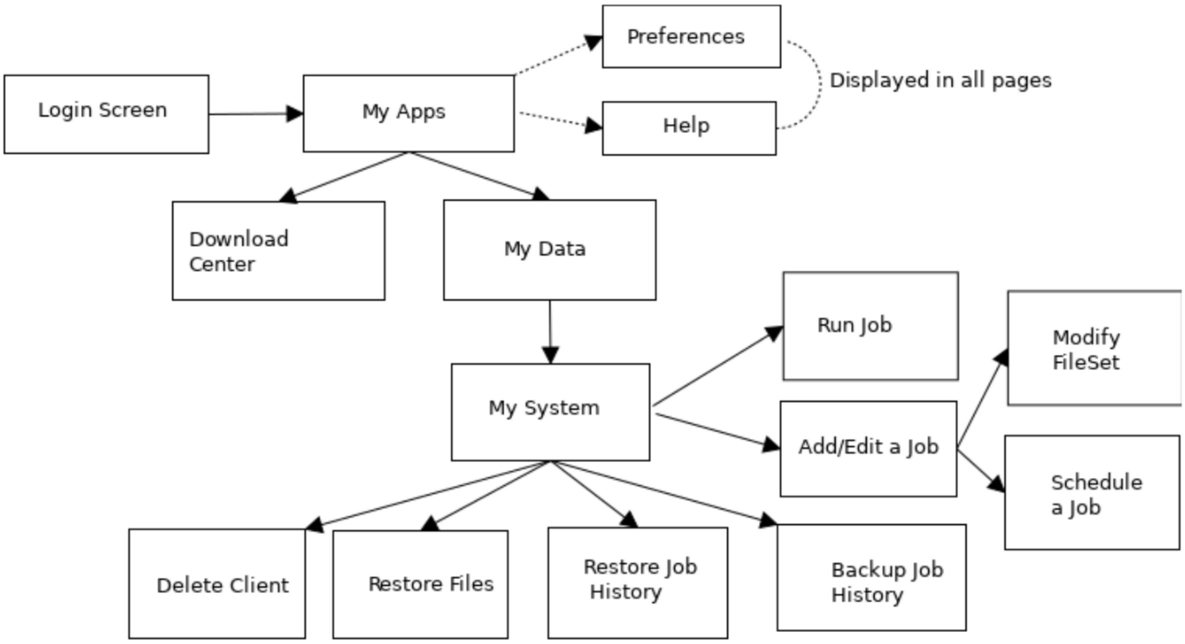


Figure 4.3: bcloud Screens Organization

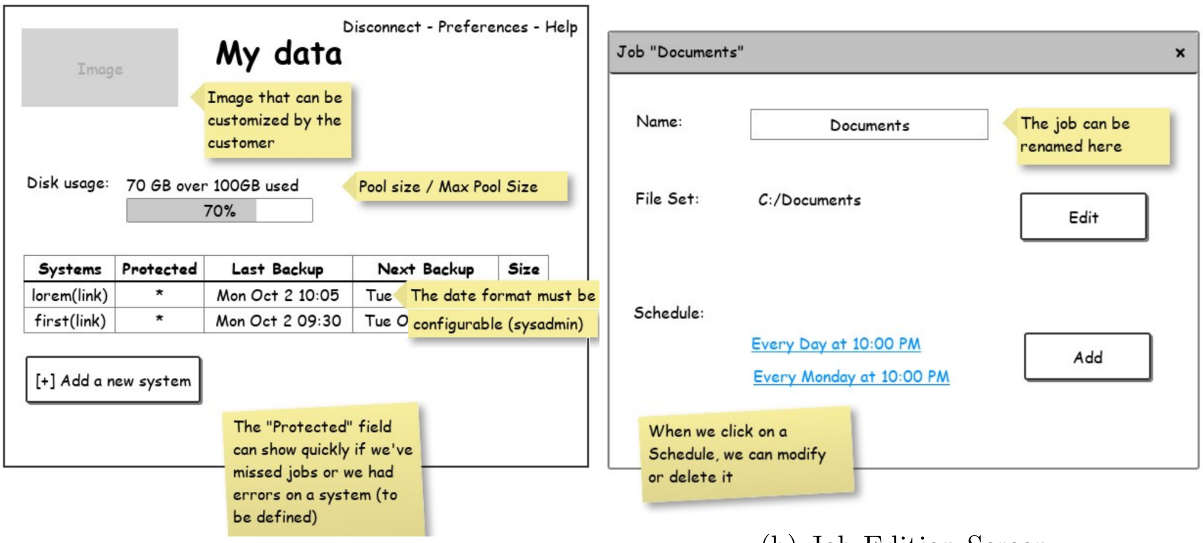
Main Screen (My Apps) The main screen will give access to the documentation, the Download Center, and the User Overview screens.

Download Center Screen As presented earlier, end users will be able to browse and download the Bacula Client package(s) for their systems.

User Overview Screen (My Data) As exhibited in Figure 4.4a, the User Overview screen will show the quota usage and the list of all Clients registered. The time of the last backup and the total job size for each Client will be displayed as well. A link to the procedure to register a new Client will be displayed. A link for each Client line will give access to the Client Overview page.

Client Overview Screen (My System) The Client Overview page will give information about the last backup job, the list of the previous Jobs, the log of each Job, the status of the previous Jobs, the time of the next Job, the Quota and percent of Quota used, the jobs that are running (number of files, number of bytes, start time, elapsed time, status), the list of all defined Jobs (FileSet content, Schedule),a link to modify the Schedule resource, a link to the Restore Interface and an option to delete the Client and all the associated data from the Director.

Job Edition Screen As viewed in the Figure 4.4b, the cloud user can create backup routines, specify folders or files for backup, and select a schedule for job automatic queue.


Disconnect - Preferences - Help

Image

My data

Image that can be customized by the customer

Disk usage: 70 GB over 100GB used

70%

Pool size / Max Pool Size

Systems	Protected	Last Backup	Next Backup	Size
lorem(link)	*	Mon Oct 2 10:05	Tue	
first(link)	*	Mon Oct 2 09:30	Tue O	

The date format must be configurable (sysadmin)

[+] Add a new system

The "Protected" field can show quickly if we've missed jobs or we had errors on a system (to be defined)

Job "Documents" ×

Name: Documents

The job can be renamed here

File Set: C:/Documents

Edit

Schedule:

Every Day at 10:00 PM

Every Monday at 10:00 PM

Add

When we click on a Schedule, we can modify or delete it

(a) Overview Screen

(b) Job Edition Screen

Figure 4.4: Mock-ups

Restore Interface Screen The Restore Interface will present the list of the Jobs, the list of the files and the end user will be able to select directories or files. Once the files are selected, the end user will be able to change the Client and the destination directory used for the restore.

Edit Schedule Screen The Edit Schedule box will offer simple options to schedule a Job, such as every day at a fixed hour; once per week at a fixed day and hour; once per month at a fixed day and hour; we might add other scheduling options in the future.

4.3 Results and Evaluation

The prototype development stage had four months of duration until bcloud Alpha version release, which involved the web interface development and backup client wizards to ease and automate their deploy.

As shown in Figure 4.5, the Main Screen (My Apps) is a clean, customizable screen that is the landing area after login. Users can choose between CSS themes, company logotype, and available applications. Newly created users at the identity service backend, such as LDAP directory services, have new tenants associated.

As displayed in Figure 4.6a, the user will access the Download Center to fetch multiplatform backup client installation packages and configuration wizards, to automatically attach its machines to the backup system. The Cloud Administrator selects these files,

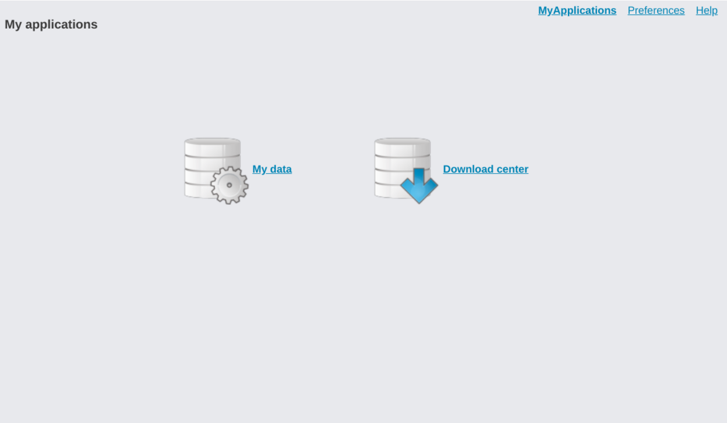
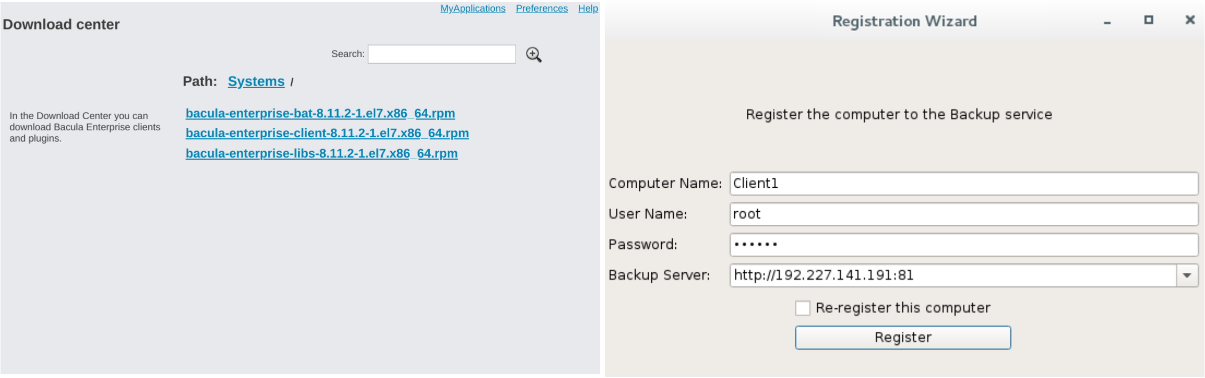


Figure 4.5: Main Screen

that is capable of restricting or providing access to specific platforms packages, application specific backup plugins, manuals and other discretionary files.

As exhibited in Figure 4.6b, the backup client Registration Wizard can run in operating systems with and without graphical interfaces. That is important especially for servers that usually doesn't have a GUI and are the principal Bacula service object.



In the Download Center you can download Bacula Enterprise clients and plugins.

bacula-enterprise-bat-8.11.2-1.el7.x86_64.rpm
bacula-enterprise-client-8.11.2-1.el7.x86_64.rpm
bacula-enterprise-libs-8.11.2-1.el7.x86_64.rpm

Register the computer to the Backup service

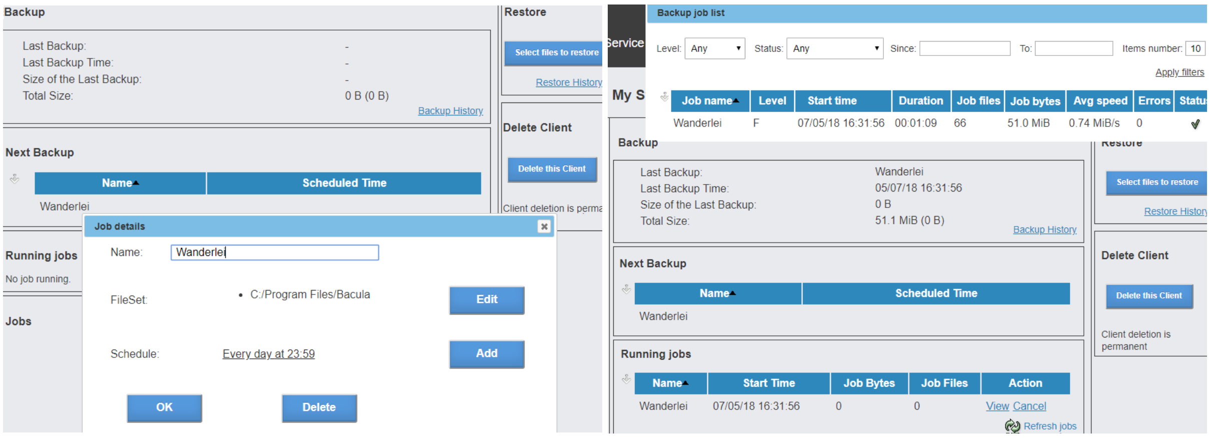
Computer Name: Client1
User Name: root
Password: ••••••
Backup Server: http://192.227.141.191:81
□ Re-register this computer

(a) (b)

Figure 4.6: bcloud Client Download Center and Configuration Wizard

Once installed, configured, and approved by the Cloud Administrator, the backup clients will appear at "My Data" bcloud screen. From this moment, as shown in Figure 4.7a, the user can proceed with backup job configuration steps, such as choosing what folders, files or applications will be backed up and schedule. Ad hoc queuing is also supported.

As presented in Figure 4.7b, a dashboard shows all terminated and on-going jobs. It displays information such as backup duration, number of files, size, transfer rate, start and termination time. The user can filter listed jobs by level, status, client, and time frame, using the same tab options.



Backup

Last Backup: -
Last Backup Time: -
Size of the Last Backup: -
Total Size: 0 B (0 B)

Client deletion is perma

Next Backup

Name	Scheduled Time
Wanderlei	

Running jobs

No job running.

Jobs

Job details

Name: Wanderlei

FileSet:
- C:/Program Files/Bacula

Schedule: Every day at 23:59

(a)

Backup job list

Level: Any Status: Any Since: To: Items number: 10

Job name	Level	Start time	Duration	Job files	Job bytes	Avg speed	Errors	Statu
Wanderlei	F	07/05/18 16:31:56	00:01:09	66	51.0 MiB	0.74 MiB/s	0	✓

Backup

Last Backup: Wanderlei
Last Backup Time: 05/07/18 16:31:56
Size of the Last Backup: 0 B
Total Size: 51.1 MiB (0 B)

Client deletion is permanent

Next Backup

Name	Scheduled Time
Wanderlei	

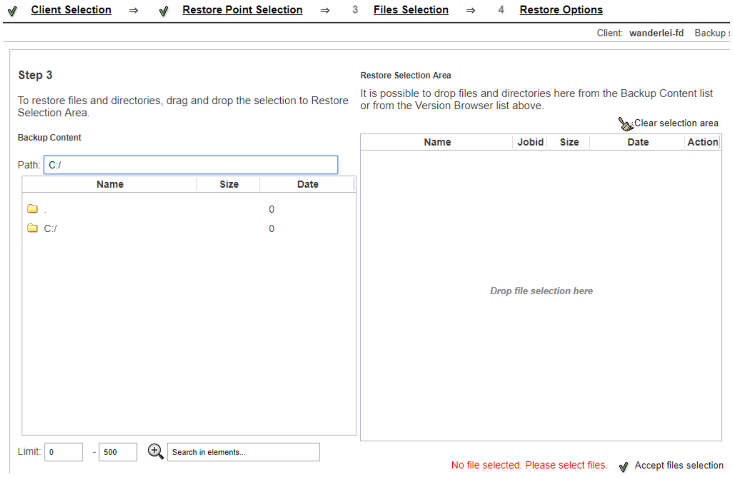
Running jobs

Name	Start Time	Job Bytes	Job Files	Action
Wanderlei	07/05/18 16:31:56	0	0	View Cancel

(b)

Figure 4.7: bcloud Job Configuration and Terminated Jobs Overview

Finally, as displayed in Figure 4.8, the Cloud User can select one or more terminated backup jobs, generally according to their termination time, and proceed with file browsing and data restore. Differential and incremental backups can be selected, manually or automatically, together with last full and other necessary jobs to provide a complete restore to a given time. All files, a full directory, a few files or even plugin backed up application data such as a database, and a single virtual machine can be selected from the selection tree.



✓ Client Selection ⇒ ✓ Restore Point Selection ⇒ 3 Files Selection ⇒ 4 Restore Options

Client: wanderlei-fd Backup :

Step 3

To restore files and directories, drag and drop the selection to Restore Selection Area.

Backup Content

Path: C:/

Name	Size	Date
.	0	
C:/	0	

Limit: 0 - 500 Search in elements...

Restore Selection Area

It is possible to drop files and directories here from the Backup Content list or from the Version Browser list above.

Name	Jobid	Size	Date	Action

Drop file selection here

No file selected. Please select files.

Figure 4.8: Restore Job File Browsing

4.4 Case Study

As exhibited in Figure 4.9, there are two essential bcloud roles: Cloud Service Administrator and the Cloud User. The first is the backup-as-a-service provider that access bcloud to perform administrative tasks or another Bacula CLI, GUI or other web-based interfaces. The second one is the backup-as-a-service customer, who receives a tenant and executes the steps to set up the backup of its data.

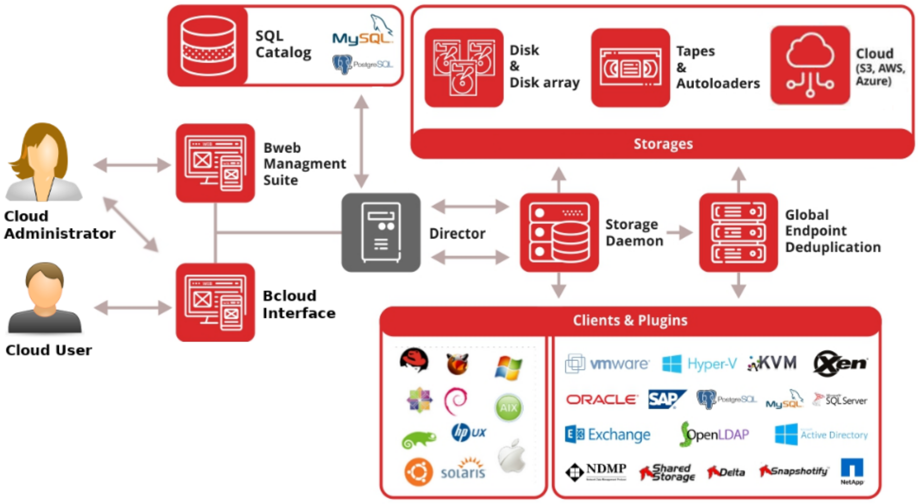


Figure 4.9: bcloud Users Roles

This case study lists a minimum workflow that each user profile will typically take in bcloud towards the BaaS configuration and a description of the prototype deploy environment, as follows.

4.4.1 Cloud Service Administrator Role

The Cloud Administrator is the internal or external service provider specialist manager, responsible for the BaaS platform availability, continuity, scalability, and performance. His roles are usually the following:

- User downloadable Bacula backup clients selection
- New user tenant in the LDAP/AD services creation
- User optional quota setup
- New backup clients registration approval
- All users backup size, status, log view and billing management
- On-going and terminated backup monitoring and reporting

Besides accessing bcloud to perform clients approval and quota management, the Cloud Administrator can also access regular Bacula interfaces, such as CLI, desktop and other web-based ones to perform traditional backup monitoring, log analysis and troubleshooting.

4.4.2 Cloud User Role

The cloud user might consist of an internal or external backup service customer, that receives tenant access, being allowed to perform one or all of the about to be mentioned tasks:

- New BaaS tenant credential reception and access
- Multi-platform Bacula backup clients download
- Client installation and registration wizard execution
- Folders and backup schedule definition
- Backup jobs cancellation and deletion
- Restore job submission

The cloud user can also receive mail notifications about performed backed up jobs and reports if configured.

4.4.3 Prototype Environment Description

To validate the product features, we set up a small environment for user tests over the Internet. The initial requirements are similar to a Bacula backup server since bcloud is just one more hosted interface. We used a machine with 64 bits virtualized CentOS operating systems with 100 GB of disk space, and 16 GB of RAM. Also, a 512 GB direct-attached solid-state drive was set up to host the global deduplication backup engine since it requires faster random access for on-the-fly processing. The backup server local network traffic relies on gigabit Ethernet connections, but the client backup traffic data goes through the Internet.

This capacity considers a small workload, estimated to support up to 100 backup clients, 50 simultaneous bcloud web interface users and 50 TB of backups. Extra backup load, especially the number of backed up files would require more resources, such as Bacula database meta-data catalog required disk space. More backed up files represent a more significant number of rows in the File path table. Nevertheless, more frequent and larger backups will also need a larger solid-state drive partition to host the growing deduplication index.

4.5 Usability Evaluation

Since the product of this work is a user interface, and even backup performance is mostly hardware and network dependent, non-functional software evaluation is out of the scope of this analysis.

As for the functional evaluation, according to Adrion et al. [100], going over a program by hand while sitting at one's desk is one of the most traditional means for analyzing a program, also called desk checking. It is also the origin of more formal techniques such as walk-throughs, inspections, and reviews. One benefit of these methods is seeing one's own errors is difficult, so it is more effective if a second party does the desk checking.

However, walk-throughs and inspections require a crew, usually directed by a moderator and including the software developer, what is not possible in the current case because of the geographically sparse developing team and has the disadvantage of a relying on a small number of users feedback.

Modernly, as stated by Brhel et al. [101], user usability testing, individual inquiry, and fast prototyping are recognized as the most frequent usability practices evaluation techniques. The authors suggest even early low fidelity prototypes and user story maps are strategic and cost-effective concepts for creating artifacts, especially for distributed teams.

Since Cloud services are user-centric, focused on user-experience and accessed by a large number of customers with different profiles and technical levels, we chose the individual inquiry technique for the bcloud evaluation in this study.

4.5.1 Sample versus Population

The user inquiry was performed using documentation, prototype user access, and an online questionnaire that contained six objective zero to ten evaluation questions. That scale was chosen for being practical and for providing a granular perception of the evaluated features. Nevertheless, a general user-feedback open question was made available, and the variety of answers were categorized and presented.

A total of 278 fillings were considered, corresponding to a population of 1000 users, with an error margin of 5% and a confidence level of 95%. We used the Normal (Gaussian or Gauss or Laplace–Gauss) continuous probability distribution, to represent real-valued random variables whose distributions are not known. The random variables observations sample averages drawn from independent distributions converge to the normal, in other words, they become normally distributed when the number of observations is sufficiently large.

The statistical population is higher than most small and medium company staff sizes, in a hypothetical and extreme scenario where every employee becomes a Backup-as-a-Service user.

Worldwide users were invited to test the bcloud interface, follow the Cloud User and Cloud Administrator workflow suggestion, and provide feedback via questionnaires. The users received a research description, the interface user manual and the credentials to access the prototype.

The poll sought heterogeneous user-profiles, with different technical background and seniority levels. Invites were sent to infrastructure, cloud, operating system, backup and even Bacula user groups. IT managers and other stakeholders from the private and government sectors, teachers, graduate and even undergraduate students from other areas such as Data Mining, Risk Management, Business, and Software Engineering also participated.

We monitored the bcloud web engine logs to make sure each invited customer was able to access the solution, perform the defined user roles and start a backup.

4.5.2 Questionnaire

We evaluated the BaaS solution according to the four cloud requirements that we chose earlier. The “Interface Reconfiguration Capabilities”, “Multi-Tenancy Support”, and “Easy New Tenants Creation” criteria had one question each. The “User-Centered Design” was divided into three questions: “Reduced User Operation Errors”, “Improved User Acceptance and Satisfaction”, “and Improved Productivity”. We present the questions as follows.

4.5.3 Interface Reconfiguration Capabilities

Knowing that the bcloud administrator can:

- Define and make available backup client installation packages, manuals and other documents specific to bcloud users
- Configure optional quotas for each user
- Change the visual themes of the solution and logo of the service provider, through CSS

Question: on a scale of zero to ten, how do you evaluate the reconfiguration, personalization and customization features of the bcloud interface?

4.5.4 Multi-Tenancy Support

As a cloud solution, bcloud adopts the user-centered privacy principle, receiving the access to a shared interface with several other users, using the same data center resource, but only is capable of seeing their own tasks and information by default.

Bacula single instance of the object code and database supports multiple customers.

Question: on a scale of zero to ten and according to your bcloud usage perception, is bcloud a multi-tenant solution?

4.5.5 Easy New Tenants Creation

Creating new tenant access to bcloud is done by integrating and creating new users in the directory service bases used by the client (e.g., LDAP and AD).

Question: on a scale of zero to ten, how easy do you consider requesting or creating new tenants for the service?

4.5.6 User-centered Design

The bcloud, as a cloud interface, aims to provide a user-centered design to attend a wide range of cloud users technical levels.

A user-centered design [102] is the use of the technology to fit human capabilities, resulting in reduced error, improved user acceptance and satisfaction, and improved productivity.

The bcloud interface access should provide a reasonable overview of the bcloud operation and user-centered qualities.

Questions:

1. On a scale of zero to ten, how do you evaluate the bcloud solution about the supposed reduced backup operation error capacities?
2. On a scale of zero to ten, how do you evaluate the solution about the hypothetic improved user acceptance and satisfaction?
3. On a scale of zero to ten, how do you evaluate the solution about the alleged improved productivity?
4. What comments and other considerations can you make about bcloud?

4.6 Answers

We present the average questionnaire evaluation for the first three questions as follows:

1. Interface reconfiguration capabilities
2. Multi-tenancy support
3. Easy new tenants creation

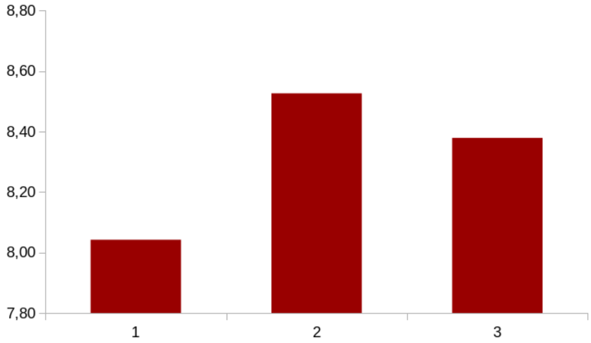


Figure 4.10: Questionnaire questions 1, 2 and 3 average evaluations.

As shown in Figure 4.10, the average replies about the investigated features deploy ranged from roughly 8 to an 8.5 score. The "Interface Reconfiguration Capabilities" was the worse evaluated criteria and the "Multi-Tenant Deploy" was the best.

We present the Design User-centered design question results as follows:

1. Reduced backup operation error capacities
2. Improved user acceptance and satisfaction
3. Improved productivity

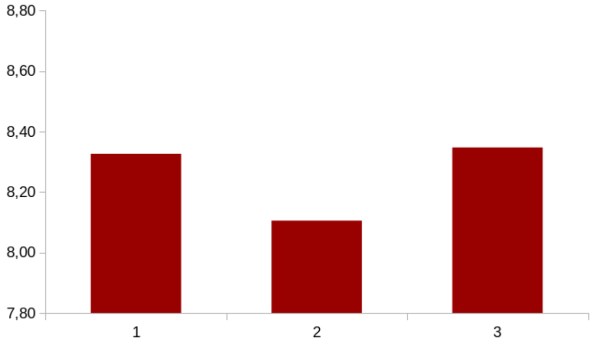


Figure 4.11: Questionnaire questions 4, 5 and 6 average evaluations.

As shown in Figure 4.11, the average replies ranged from 8.1 to an 8.35 score. The "Improved User Acceptance and Satisfaction" interface quality was the worse evaluated, and the "Reduced Backup Operation Error Capabilities" was the best-evaluated criteria.

The user-feedback open question (7) received a high number of different themes, and they were summarized as follows:

1. Usability and user technical level concerns
2. Safety and performance considerations
3. More User experience work is needed
4. Innovative solution
5. Useful software
6. Adherent to the cloud trend

The three first topics were considered more detrimental to the user interface perception and will be taken into consideration for the future prototype improvement. The last three are more positive replies which reaffirm the importance of the Backup-as-a-Service solution development.

4.7 Result Analysis

The overall average objective questions mark was 8.29%, what indicates a very satisfactory perception of the bcloud BaaS interface prototype in such a short time-frame. The prototype construction, evaluation, and user-feedback were essential to investigate the new backup multi-tenant self-service paradigm potential, challenges, and possibilities.

We present and comment each one of the objective user-inquiry answers result from highest to the lowest as follows.

Multi-Tenancy Support - 8.52. This result is a little bit surprising, because there are no gray shades of multi-tenancy on the prototype deployment. By default, each user only sees and manages its own backup clients and jobs, even though bcloud has multi-party capabilities. We report the score not being perfect to the user lack of knowledge about multi-tenancy, and to a fail in explain that to the inquired subjects.

Easy new tenants creation - 8.4. We chose not to create an "add user" feature directly on the bcloud interface, because it would elevate the complexity considerably. There are a number of different identity bases solutions, such as Open LDAP, Microsoft Active Directory and every public cloud provider also have their own. It is effort wise to abstract this feature, and let the cloud provider use the native directory server tools for user creation. Nevertheless, we consider the feedback very satisfactory.

Improved productivity - 8.37. As mentioned on the literature review, many backup administrators just want to use the backup software defaults. We associate that fact to the good grade achieved in this question, and automated features such as the backup clients configuration must have contributed.

Reduced backup operation error capacities - 8.32. The bcloud interface abstracts many of the complexities of current data center backup software ones. The users can select and download only the backup clients and plug-ins that the administrator provides, and the installation wizard makes much easier to install the new clients. The user-experience focused design abstracts complicated backup elements such as storage capacity, type, pools and volumes. Users of virtually any level can operate bcloud.

Improved user acceptance and satisfaction - 8.07. This result was a little disappointing, but it can be explained on the lack of demonstrations, any "help" content or built-in tutorial, due to the lack of time. Even being easier, some users with no backup knowledge might have struggled to understand some of the features.

Interface Reconfiguration Capabilities - 8.04. We report this lower result to the fact there was no mention on the bcloud user manual on how to customize the bcloud interface, even though it is CSS based. CSS interfaces might be easily edited in order to match the Cloud Provider visual identity. We are going to address this issue with a customization manual and perhaps some sort of built-in editor.

Since we deployed the most critical BaaS features, it is now easier to enhance details such as improved user acceptance, reconfiguration capabilities, and to expand bcloud with all prior listed Cloud requirements. Nevertheless, cloud providers can safely use bcloud in production environments, where new users and tests will provide further feedback.

4.8 Chapter Summary

This chapter described the prototype development, deployed stages, and details most crucial interface screens. A case study explains the principal user roles and the prototype deploy environment. The evaluation methodology, sample definition and the user questionnaire and the answers were summarized and presented. The results indicated that our decisions on this project were adequate and that the evaluation was very satisfactory. The next chapter draws the study conclusions and future work.

Chapter 5

Conclusion

Backup is the replica of any data that can be used to restore its original form and is often stored to lower cost of high capacity removable media, such as magnetic tapes, that are stored in fireproof safes or another protected physical environment. However, developers should update old traditional backup software to the new Cloud technologies such as Object Storage and user-centered design.

Cloud computing is a long-held idea of computing as a general utility. It promises to shift data and computational services from individual devices to distributed architectures with a general cost reduction for disaster recovery mechanisms and inherent data geographical distribution.

The total amount of digital data created worldwide more than doubles every two years. It is growing from 4.4 zettabytes in 2013 to 44 zettabytes by 2020. The considerable increase in data generation and processing confers unprecedented pressure to backup environments. Data itself has become a valuable asset, and it is more important to protect it using highly reliable backup systems. The Recovery Time Objective (RTO) and Recovery Point Objective (RPO) are still the primary backup parameters that must be observed and kept at optimal values.

There are also a variety of other cloud computing development challenges, such as risk management, trust and recovery mechanisms should be studied to provide business continuity and better user satisfaction.

Backup routines will still be relevant with growing Cloud adoption, and there are many useful Cloud backup architectures, such as Remote Backup to the Cloud, Local Backup from the Cloud, Cloud Geographical Redundancy and Backup (GRB), Inter-private Cloud Storage (IPCS) and Secure-Distributed Data Backup (SDDB).

The ideal modern currently backup and recovery software products should not only provide features to attend a traditional data center but allow the integration and exploration of the growing Cloud, including "backup client as a service" and "backup storage as a service". A data-center Backup-as-a-Service model can bring benefits to both Cloud Providers and Cloud Users.

The present study developed and prototyped a Backup as a Service (BaaS) solution, under the "Remote Backup to the Cloud" architecture. The cloud/backup parameters, cloud backup challenges, researched architectures and BaaS system requirements were determined. A set of features were selected to be developed and implemented to attend the chosen architecture.

There are many Cloud Services Provider macro-requirements for BaaS, such as autonomy, cloud scalability, self-description, fault tolerance, interoperability, load balancing, multi-party, multi-tenancy, optimal provisioned infrastructure usage, quality of service, standard interfaces, storage, data and workload management. As for cloud user requirements, we surveyed the automation, backup standardization, data consistency, data integrity, NAT and firewall transversing capabilities, non-disruptive backup, SLA compliance, user-centric privacy, consumption-based billing and user experience focused design.

We surveyed backup software alternatives according to consultancy indicated leaders and the most popular solutions. According to the BaaS macro-requirements and cost, we chose Bacula's Enterprise version for the BaaS Interface prototype deployment. The already existing REST API feature, open catalog format, cloud backup storage capabilities, a variety of specific applications backup plug-ins and reasonable cost were the main reasons for this decision.

However, the usage of the REST API makes bcloud likely to work with the Bacula open source version and even other backup softwares that might happen to have that cloud industry feature, even if some command changes are needed. In our opinion, that adds value to the proposal.

We defined a list of 17 requirements for an ideal BaaS interface, but for the time restrictions, we prioritized and evaluated the four feasible within the limited time frame. They were: reconfiguration personalization and customization capabilities; multi-tenancy support; easy new tenants creation; and user-centered design. Considering these features, we developed a new web interface named bcloud.

We presented the BaaS topology solution and the screen sketches. The prototype development stage had four months of duration until the Alpha bcloud BaaS interface version release, which involved the web system development and backup client wizards to ease and automate their deploy. We also exhibited the final prototype screens in Chapter 4.

We built a Case Study and an on-line prototype evaluation scenario. We explained the evaluation methodology selection criteria and chose the user inquiry through Internet questionnaires to validate and evaluate the requirements deploy. As the result of this work, the users received a research description, the interface user manual and the credentials to access the prototype interface.

A sample was determined to attend to most Small and Medium Business in an extreme scenario where every corporate user would be a BaaS user, rendering a total of 278 fillings. That corresponds to a thousand users, with an error margin of 5% and a confidence level of 95%.

The overall average objective zero to ten questions mark was 8.29%, indicating a very satisfactory perception of the bcloud BaaS interface prototype in such a short time-frame.

The "Multi-Tenancy Support" and "Easy new tenants creation" were the best evaluated criteria according to the user feedback. The "Improved User Acceptance and Satisfaction" quality and the "Interface Reconfiguration Capabilities" were the worse evaluated topics, and we are going to improve the interface according to these feedbacks.

The user-feedback open question received a high number of different themes, and they were summarized as follows: "usability and user technical level concerns", "safety and performance considerations", and "more user experience work is needed", that were interpreted as more negative or critic observations and we are going to take them in consideration for the future prototype improvement. We received the "innovative solution", "useful software", and "cloud trend adherent" categories as common positive remarks that reaffirm the BaaS development opportunity and encourages the prototype evolution.

As a prototype and according to the user feedback, bcloud accomplished its purpose of being considered as a potential Cloud product, although we should improve the already deployed features and develop other non-deployed requirements.

The adoption of a BaaS model should make backup more flexible, popular and inexpensive, during times when the press frequently notices data-loss disasters, cyber-terrorist activities are increasing and the companies pressures for lower cost solutions.

5.1 Future Works

As an improvement to the prototype deployment, we are going to add interface help content and demonstrations, so even most unexperienced user will be able to use the interface. That might improve the end-user acceptance and satisfaction perception.

Still, we plan to incorporate a CSS code editor to the bcloud administrator interface, so it will be easier for the administrator to perform design configuration changes. Today there are many Internet browser extensions such as Stylebot [103], which allows to easily modify and preview websites appearance on-the-fly, generate the code automatically, copy and apply to the bcloud editor.

As mentioned in the study, Bacula has a client-initiated connection backup mechanism but this limits the backup system central management and operation. The bcloud installation wizard might be improved to automate a VPN client configuration for bcloud service or another technique that will involve Bacula source code modifications.

We are going to assist, study and evaluate a bcloud production deployment, and will conduct another inquiry to verify the product evolution.

Bibliography

[1] M. R. Raju, J. P. Prakash, and G. R. Rao, “Disaster Recovery of Servers using Virtualized Cloud Computing,” 2012. [Online]. Available: http://www.conference.bonfring.org/papers/cjits_icarmmiem2014/mmiem-17.pdf vii, 7, 8

[2] Y. Ueno, N. Miyaho, S. Suzuki, and K. Ichihara, “Performance Evaluation of a Disaster Recovery System and Practical Network System Applications.” IEEE, Aug. 2010, pp. 195–200. [Online]. Available: http://ieeexplore.ieee.org/document/5635011/ vii, 16, 17, 40

[3] D. Russell, P. Rinnen, and R. Rhame, “Magic Quadrant for Data Center Backup and Recovery Software,” 2016. vii, 1, 2, 25, 26, 28, 29, 30, 44

[4] Google, “Google Trends,” 2017. [Online]. Available: /trends/explore vii, 26

[5] C. Preston, *Backup and Recovery: Inexpensive Backup Solutions for Open Systems.* "O’Reilly Media, Inc.", 2007, google-Books-ID: M9mbAgAAQBAJ. vii, 10, 22, 33

[6] “frapcom Wikimedia,” 2015. [Online]. Available: http://www.frapcom.be/mediawiki/index.php?title=Bacula vii, 34

[7] P. d. Guise, *Enterprise Systems Backup and Recovery: A Corporate Insurance Policy.* USA: CRC Press, 2008, google-Books-ID: 2OtqvySBTu4C. 1, 37, 38, 40

[8] IDC, “Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives | The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.” [Online]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm 1

[9] D. Russel, “The Future of Backup May Not Be Backup,” 2011. 1, 44

[10] T. J. Silva, “Uma arquitetura de cloud storage para backup de arquivos,” 2014. [Online]. Available: http://repositorio.ufpe.br/handle/123456789/18013 1

[11] J. Kaiser, T. Süß, L. Nagel, and A. Brinkmann, “Sorted deduplication: How to process thousands of backup streams,” in *Mass Storage Systems and Technologies (MSST), 2016 32th Symposium on. IEEE*, 2016. [Online]. Available: http://storageconference.us/2016/Papers/SortedDeduplication.pdf 1

[12] M. A. Khoshkholghi, A. Abdullah, R. Latip, S. Subramaniam, and M. Othman, “Disaster Recovery in Cloud Computing: A Survey,” *Computer and Information Science*, vol. 7, no. 4, p. 39, Sep. 2014. [Online]. Available: http://www.ccsenet.org/journal/index.php/cis/article/view/37067 2, 4, 7, 8, 9, 10, 11

[13] H. Faria, *Bacula The Open Source Backup Software*, Dec. 2016. 3, 36

[14] H. Faria, R. Carvalho, and P. Solis, “Storage Growing Forecast with Bacula Backup Software Catalog Data Mining,” in *Computer Science & Information Technology (CS & IT).* Academy & Industry Research Collaboration Center (AIRCC), Mar. 2017, pp. 185–196. [Online]. Available: http://airccj.org/CSCP/vol7/csit76618.pdf 3, 26

[15] H. Faria, J. Luiz Bordim, and P. Solis Barreto, “Backup Storage Block Level Deduplication with DDUMBFS and BACULA,” *International Journal of Advanced Information Technology*, vol. 7, no. 4, pp. 1–9, Aug. 2017. [Online]. Available: http://aircconline.com/ijait/V7N4/7417ijait01.pdf 3

[16] H. Faria, “A Hadoop Open Source Backup Solution,” 2018. 3

[17] M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, and A. Rabkin, “A view of cloud computing,” *Communications of the ACM*, vol. 53, no. 4, p. 50, Apr. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1721654.1721672 4

[18] R. Buyya, J. Broberg, and A. Gościnski, *Cloud computing: principles and paradigms.* Hoboken, N.J: Wiley, 2011, oCLC: ocn606774387. 4

[19] L. Columbus, “Roundup Of Cloud Computing Forecasts,” 2017. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2017/04/29/roundup-of-cloud-computing-forecasts-2017/#2f36374d31e8 4

[20] S. Ried, H. Kisker, and A. Bartels, “Sizing The Cloud,” 2011. [Online]. Available: https://www.forrester.com/report/Sizing+The+Cloud/-/E-RES58161 4

[21] O. Arean, “Disaster recovery in the cloud,” *Network Security*, vol. 2013, no. 9, pp. 5–7, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1353485813701016 4

[22] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” 2011. 5, 6

[23] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, “NIST Cloud Computing Reference Architecture.” IEEE, Jul. 2011, pp. 594–596. [Online]. Available: http://ieeexplore.ieee.org/document/6012797/ 6

[24] “IBM United States.” [Online]. Available: https://www.ibm.com/us-en/ 7

[25] O. Alhazmi and Y. Malaiya, “Evaluating Disaster Recovery Plans Using the Cloud,” 2013. 8

[26] T. Wood, E. Cecchet, K. K. Ramakrishnan, P. J. Shenoy, J. E. van der Merwe, and A. Venkataramani, "Disaster Recovery as a Cloud Service: Economic Benefits & Deployment Challenges." *HotCloud*, vol. 10, pp. 8–15, 2010. [Online]. Available: https://www.usenix.org/legacy/events/hotcloud10/tech/full_papers/Wood.pdf 9

[27] H. Kashiwazaki, "Practical uses of cloud computing services in a Japanese university of the arts against aftermath of the 2011 Tohoku earthquake," in *Proceedings of the 40th annual ACM SIGUCCS conference on User services.* ACM, 2012, pp. 49–52. [Online]. Available: http://dl.acm.org/citation.cfm?id=2382467 9

[28] S. Prakash, S. Mody, A. Wahab, S. Swaminathan, and R. Paramount, "Disaster recovery services in the cloud for SMEs," in *Cloud Computing Technologies, Applications and Management (ICCCTAM), 2012 International Conference on.* IEEE, 2012, pp. 139–144. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6488087/ 10

[29] V. Javaraiah, "Backup for cloud and disaster recovery for consumers and SMBs," in *Advanced Networks and Telecommunication Systems (ANTS), 2011 IEEE 5th International Conference on.* IEEE, 2011, pp. 1–3. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6163671/ 10, 13

[30] O. Alhazmi and Y. Malaiya, "Assessing Disaster Recovery Alternatives: On-site, Colocation or Cloud," 2012. 10

[31] F. Sabahi, "Cloud computing security threats and responses," in *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on.* IEEE, 2011, pp. 245–249. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6014715/ 10

[32] I. P. T. News, "FBI: Cyber Terrorism is Growing Threat," 2010. [Online]. Available: http://www.investigativeproject.org/1831/fbi-cyber-terrorism-is-growing-threat 10

[33] G. O'Gorman and G. McDonald, *Ransomware: A growing menace.* Symantec Corporation, 2012. [Online]. Available: http://www.01net.it/whitepaper_library/Symantec_Ransomware_Growing_Menace.pdf 10

[34] M. Ji, A. C. Veitch, J. Wilkes, and others, "Seneca: remote mirroring done write." in *USENIX Annual Technical Conference, General Track*, 2003, pp. 253–268. [Online]. Available: http://static.usenix.org/events/usenix03/tech/full_papers/full_papers/ji/ji.pdf 10

[35] Z. Jian-hua and Z. Nan, "Cloud Computing-based Data Storage and Disaster Recovery." IEEE, Aug. 2011, pp. 629–632. [Online]. Available: http://ieeexplore.ieee.org/document/6041774/ 11, 15

[36] M. Pokharel, S. Lee, and J. S. Park, "Disaster Recovery for System Architecture Using Cloud Computing." IEEE, Jul. 2010, pp. 304–307. [Online]. Available: http://ieeexplore.ieee.org/document/5598055/ 11, 14

[37] Y. Hua, X. Liu, and D. Feng, "Cost-Efficient Remote Backup Services for Enterprise Clouds," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 5, pp. 1650–1657, Oct. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7435279/ 11

[38] J. Xu, W. Zhang, S. Ye, J. Wei, and T. Huang, "A Lightweight Virtual Machine Image Deduplication Backup Approach in Cloud Environment." IEEE, Jul. 2014, pp. 503–508. [Online]. Available: http://ieeexplore.ieee.org/document/6899254/ 11, 22

[39] H. Camacho, A. Brambila, A. Peña, and J. Vargas, "A Cloud Environment for Backup and Data Storage," 2014. 12

[40] H. Mao, N. Xiao, Y. Lu, and H. Xu, "Somersault cloud: Toward a cloud-of-clouds service for personal backup," in *Computing, Networking and Communications (ICNC), 2013 International Conference on.* IEEE, 2013, pp. 461–464. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6504128/ 12

[41] B. I. Ismail, M. N. M. Mydin, and M. F. Khalid, "Architecture of scalable backup service for private cloud," in *Open Systems (ICOS), 2013 IEEE Conference on.* IEEE, 2013, pp. 174–179. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/6735069/ 13, 16, 17, 21, 22, 23

[42] J. I. Khan and O. Y. Tahboub, "Peer-to-Peer Enterprise Data Backup over a Ren Cloud." IEEE, Apr. 2011, pp. 959–964. [Online]. Available: http://ieeexplore.ieee.org/document/5945364/ 14

[43] B. P. Rimal, A. Jukan, D. Katsaros, and Y. Goeleven, "Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach," *Journal of Grid Computing*, vol. 9, no. 1, pp. 3–26, Mar. 2011. [Online]. Available: http://link.springer.com/10.1007/s10723-010-9171-y 17, 18, 19, 20, 21, 23, 24

[44] V. Werner, "A Word on Scalability - All Things Distributed," 2006. [Online]. Available: http://www.allthingsdistributed.com/2006/03/a_word_on_scalability.html 18

[45] "Web Services Architecture." [Online]. Available: https://www.w3.org/TR/2004/NOTE-ws-arch-20040211/relwwwrest 20

[46] A. Cavoukian, "Privacy and Digital Identity: Implications for the Internet," 2008. 21, 23

[47] "Using VMware Infrastructure for Backup and Restore," 2006. 22

[48] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, "Extreme binning: Scalable, parallel deduplication for chunk-based file backup," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems, 2009. MASCOTS'09. IEEE International Symposium on.* IEEE, 2009, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/5366623/ 22

[49] A. Chervenak, V. Vellanki, and Z. Kurmas, “Protecting file systems: A survey of backup techniques,” in *Joint NASA and IEEE Mass Storage Conference*, vol. 99, 1998. [Online]. Available: http://www.storageconference.us/1998/papers/a1-2-CHERVE.pdf 22

[50] M. Stiemerling, “NAT and firewall traversal issues of host identity protocol (HIP) communication,” 2008. [Online]. Available: https://tools.ietf.org/html/rfc5207.txt 22

[51] W. Louth, “Metering the Cloud: Applying Activity Based Costing (ABC) from Code Profiling up to Performance Cost Management of Cloud Computing,” 2009. 24

[52] Arcserve, “Backup and recover critical business applications,” Jul. 2015. [Online]. Available: https://arcserve.com/data-protection-solutions/backup-business-applications/ 26

[53] “Best open source network data backup and recovery software,” 2017. [Online]. Available: https://www.baculasystems.com/ 26, 27

[54] “Technology Research | Gartner Inc.” [Online]. Available: http://www.gartner.com/technology/home.jsp 26

[55] “Data backup tools from Bacula Systems.” [Online]. Available: https://www.baculasystems.com/products/bacula-enterprise-data-backup-tools 27

[56] “SQL Server 2016 Microsoft.” [Online]. Available: https://www.microsoft.com/pt-br/sql-server/sql-server-2016 27

[57] “Server Virtualization with VMware vSphere.” [Online]. Available: https://www.vmware.com/products/vsphere.html 27

[58] “KVM.” [Online]. Available: https://www.linux-kvm.org/page/Main_Page 27

[59] “The worlds leading software development platform GitHub.” [Online]. Available: https://github.com/ 27

[60] “Bareos Backup Archiving REcovery Open Sourced Main Reference.” [Online]. Available: http://doc.bareos.org/master/html/bareos-manual-main-reference.html 27

[61] “Pricing - Bareos.” [Online]. Available: https://www.bareos.com/en/Pricing.html 27

[62] “Commvault - The Award-Winning Data Management Platform For Enterprises.” [Online]. Available: https://www.commvault.com/ 27

[63] “Cloud Backup and Recovery - Popular Bundles.” [Online]. Available: https://www.commvault.com/ 28

[64] Commvault, "Data Protection and Recovery Agents," 2018. [Online]. Available: http://documentation.commvault.com/commvault/v11/article?p=landing_pages/c_data_protection.htm 28

[65] EMC, "Dedicated Networks for IP Storage," 2015. 28

[66] ——, "EMC Avamar Data Store," 2012. [Online]. Available: https://www.emc.com/collateral/software/data-sheet/h3454-avamar-data-stores.pdf 28

[67] ——, "EMC Community Network - DECN: Concurrent sessions / backup !!" 2014. [Online]. Available: https://community.emc.com/thread/191904?start=0&tstart=0 28

[68] IBM, "IBM Spectrum Protect," Aug. 2017. [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=TSD03066USEN 29

[69] ——, "Price Change(s):Price Changes: IBM Spectrum Scale Price," Oct. 2015. [Online]. Available: //www.ibm.com/products/us/en/ 29

[70] S. Oehme, "IBM Spectrum Scale Performance and sizing update," p. 36, 2015. 29

[71] "Veeam Backup and Replication - VM backup, restore, replication." [Online]. Available: https://www.veeam.com/vm-backup-recovery-replication-software.html 29

[72] "Veeam Availability for the Always-On Enterprise." [Online]. Available: https://www.veeam.com/ 29

[73] Microsoft, "Active Directory Sites and Services." [Online]. Available: https://technet.microsoft.com/en-us/library/cc730868(v=ws.11).aspx 29

[74] ——, "Secure Enterprise Email Solutions for Business Exchange," 2017. [Online]. Available: https://products.office.com/en-gb/exchange/email 29

[75] "Server Virtualization—Windows Server 2016 | Microsoft." [Online]. Available: https://www.microsoft.com/en-us/cloud-platform/server-virtualization 29

[76] "Buy Veeam Backup Essentials and Pricing and Packaging." [Online]. Available: https://www.veeam.com/veeam-backup-essentials-pricing.html 29

[77] Veeam, "Sizing and System Requirements - Veeam Backup & Replication Best Practices." [Online]. Available: https://bp.veeam.expert/architecture-overview/sizing-and-system-requirements 30

[78] "NetBackup." [Online]. Available: https://www.veritas.com/product/backup-and-recovery/netbackup-8 30

[79] "Database 12c | Oracle." [Online]. Available: https://www.oracle.com/database/index.html 30

[80] A. Scott, "NetBackup 7 Pricing Licensing Overview." 30

[81] Bacula, "Bacula Frequently Asked Questions," 2015. [Online]. Available: http://www.bacula.org/7.2.x-manuals/en/problems/Bacula_Frequently_Asked_Que.html 32

[82] X. Zhang, Z. Tan, and S. Fan, "NSBS: Design of a Network Storage Backup System," 2015. 32

[83] "Best open source network data backup and recovery software." [Online]. Available: https://www.baculasystems.com/ 32

[84] "Google Trends: Arcserve, Bacula, Netbackup popularity search worldwide," 2016, 1. [Online]. Available: https://www.google.com.br/trends/explore?date=all&q=arcserve,bacula,netbackup 32

[85] K. Sibbald, "Main Reference," 2011. [Online]. Available: http://www.bacula.org/7.4.x-manuals/en/main/Main_Reference.html 32

[86] ——, "Bacula Problem Resolution Guide," Aug. 2015. [Online]. Available: http://www.bacula.org/7.2.x-manuals/en/problems/Problem_Resolution_Guide.html 32

[87] "MySQL." [Online]. Available: https://www.mysql.com/ 32

[88] "PostgreSQL: The world's most advanced open source database." [Online]. Available: https://www.postgresql.org/ 32

[89] X. Zhang, Z. Tan, and S. Fan, "NSBS: Design of a Network Storage Backup System," *World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, vol. 9, no. 11, pp. 1245–1254, 2015. [Online]. Available: http://www.waset.org/publications/10002702 33

[90] K. Sibbald, "Creating a universally deduplicatable archive volume," Patent, 2016, classificação internacional G06F17/30; Classificação cooperativa G06F17/30156, G06F17/30073, G06F17/30194. [Online]. Available: http://www.google.com/patents/US20160055169 33

[91] W. Xia, H. Jiang, D. Feng, and L. Tian, "Combining Deduplication and Delta Compression to Achieve Low-Overhead Data Reduction on Backup Datasets." IEEE, Mar. 2014, pp. 203–212. [Online]. Available: http://ieeexplore.ieee.org/document/6824428/ 34

[92] K. Sibbald, "Main Reference Manual," 2017. 35

[93] Bacula Systems, "Corporate data backup software features from Bacula Systems." [Online]. Available: https://www.baculasystems.com/products/bacula-enterprise/features 35

[94] R. Pushan and D. Russel, "Challenging Common Practices for Backup Retention," Gartner, Inc., USA, Tech. Rep. G00278794, Jul. 2015. 39

[95] LZO, "LZO real time compression library." [Online]. Available: http://www.oberhumer.com/opensource/lzo/ 39

[96] GZIP, "The gzip home page." [Online]. Available: http://www.gzip.org/ 39

[97] "OpenLDAP, Main Page," 2014. [Online]. Available: https://www.openldap.org/ 41

[98] K. Heslin, "2014 Data Center Industry Survey," Nov. 2014. [Online]. Available: https://journal.uptimeinstitute.com/2014-data-center-industry-survey/ 43

[99] G. Amvrosiadis and M. Bhadkamkar, "Identifying Trends in Enterprise Data Protection Systems." in *USENIX Annual Technical Conference*, 2015, pp. 151–164. [Online]. Available: https://www.usenix.org/sites/default/files/conference/protected-files/atc15_slides_amvrosiadis.pdf 43

[100] W. R. Adrion, M. A. Brasntad, and J. C. Cherniavsky, "Validation, Verification, and Testing of Computer Software," p. 34, 1982. 56

[101] M. Brhel, H. Meth, A. Maedche, and K. Werder, "Exploring principles of user-centered agile software development: A literature review," *Information and Software Technology*, vol. 61, pp. 163–181, May 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950584915000129 56

[102] M. R. Endsley, *Designing for Situation Awareness: An Approach to User-Centered Design, Second Edition.* CRC Press, Apr. 2016, google-Books-ID: eRPBkapAsggC. 58

[103] Stylebot, "Stylebot," 2013. [Online]. Available: https://chrome.google.com/webstore/detail/stylebot/oiaejidbmkiecgbjeifoejpgmdaleoha 64