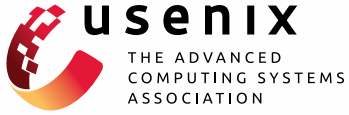


Identifying Trends in Enterprise Data Protection Systems

George Amvrosiadis, ***University of Toronto;*** **Medha Bhadkamkar,** ***Symantec Research Labs***

https://www.usenix.org/conference/atc15/technical-session/presentation/amvrosiadis

This paper is included in the Proceedings of the 2015 USENIX Annual Technical Conference (USENIC ATC '15).

July 8–10, 2015 • Santa Clara, CA, USA

ISBN 978-1-931971-225

Open access to the Proceedings of the 2015 USENIX Annual Technical Conference (USENIX ATC '15) is sponsored by USENIX.

Identifying Trends in Enterprise Data Protection Systems

George Amvrosiadis
Dept. of Computer Science, University of Toronto
gamvrosi@cs.toronto.edu

Medha Bhadkamkar
Symantec Research Labs
medha_bhadkamkar@symantec.com

Abstract

Enterprises routinely use data protection techniques to achieve business continuity in the event of failures. To ensure that backup and recovery goals are met in the face of the steep data growth rates of modern workloads, data protection systems need to constantly evolve. Recent studies show that these systems routinely miss their goals today. However, there is little work in the literature to understand why this is the case.

In this paper, we present a study of 40,000 enterprise data protection systems deploying Symantec NetBackup, a commercial backup product. In total, we analyze over a million weekly reports which have been collected over a period of three years. We discover that the main reason behind inefficiencies in data protection systems is misconfigurations. Furthermore, our analysis shows that these systems grow in bursts, leaving clients unprotected at times, and are often configured using the default parameter values. As a result, we believe there is potential in developing automated, self-healing data protection systems that achieve higher efficiency standards. To aid researchers in the development of such systems, we use our dataset to identify trends characterizing data protection systems with regards to configuration, job scheduling, and data growth.

1 Introduction

Studies analyzing the characteristics of storage systems are an important aid in the design and implementation of techniques that can improve the performance and robustness of these systems. In the past 30 years, numerous file system studies have investigated different aspects of desktop and enterprise systems [2, 6, 7, 19, 30, 39, 47, 51, 55, 56]. However, little work has been published to provide insight in the characteristics of backup systems, focusing on deduplication rates [52], and the characteristics of the file systems storing the backup images [66]. With this study, we look into the backup application generating these images, their internal structure, and the characteristics of the jobs that created them.

Modern data growth rates and shorter recovery windows are driving the need for innovation in the area of data protection. Recent surveys of CIOs and IT professionals indicate that 90% of businesses use more than two backup products [18], and only 28% of backup jobs complete within their scheduled window [34, 65]. The goal of this study is to investigate how data protection systems are configured and operate. Our analysis shows that the inefficiency of backup systems is largely attributed to misconfigurations. We believe automating configuration management can help alleviate these configuration issues significantly. Our findings motivate and support research on automated data protection [22, 27], by identifying trends in data protection systems, and related directions for future research.

Our study is based on a million weekly reports collected in a span of three years, from $40,000$ enterprise backup systems, also referred to as *domains* in the rest of the paper. Each domain is a multi-tiered network of backup servers deploying Symantec NetBackup [61], an enterprise backup product. To the best of our knowledge, this dataset is the largest in existing literature in terms of both the number of domains, and the time span covered. As a result, we are able to analyze the characteristics of a diverse domain population, and its evolution over time.

First, we investigate how backup domains are configured. Identifying common growth trends is useful for provisioning system resources, such as network or storage bandwidth, to accommodate future growth. We find that the population of protected client machines grows in bursts and rarely shrinks. Furthermore, domains protect data of a single type, such as database files or virtual machines, regardless of domain size. Overall, our findings suggest that automated configuration is an important and feasible direction for future research to accommodate growth bursts in the number of protected clients.

The configuration of a backup system, with regards to job frequency and scheduling, is also an important contributor to resource consumption. Understanding common practices employed by systems in the field can give us better insight in the load that these systems face, and the characteristics of that load. To derive these trends, we analyzed 210 million jobs performing a variety of tasks, ranging from data backup and recovery, to management

Characteristic	Observation	Section	Previous work
System setup	The initial configuration period of backup domains is at least 3 weeks.	4.1	None
Protected clients	Clients tend to be added to a domain in groups, on a monthly basis.	4.2	None
Backup policies	82% of backup domains protect one type of data.	4.3	None
	The number of backup job policies in a domain remains mostly fixed. Also, 79% of clients subscribe to a single policy.	4.4	None
Job frequency	Full backups tend to occur every few days, while incremental ones occur daily. Recovery operations occur for few domains, on a weekly or monthly basis.	5.2	None
	Users prefer default scheduling windows during weekdays, resulting in nightly bursts of activity.	5.3	None
Job sizes	Incremental and full backups tend to be similar to each other in terms of size and number of files. Recovery jobs restore either few files and bytes, or entire volumes.	6.1	Considers file sizes instead [66]
Deduplication ratios	Deduplication can result in the reduction of backup image sizes by more than 88%, despite average job sizes ranging in the tens of gigabytes.	6.2	We confirm their findings [66]
Data retention	Incremental backups are retained for weeks, while full backups are retained for months and retention depends on their scheduling frequency.	6.3	We confirm their findings [66]

Table 1: A summary of the most important observations of our study.

of backup archives. We find that jobs occur in bursts, due to the preference of default scheduling parameters by users. Moreover, job types are strongly correlated to specific days and times of the week. To avoid these bursts of activity, we expect future backup systems to follow more flexible scheduling plans based on data protection guarantees and resource availability [4, 26, 48].

Finally, successful resource provisioning for backup storage capacity requires data growth rate knowledge. Our results show that jobs in the order of tens of GBs are the norm, even with deduplication ratios of 88%. Also, retention periods for these jobs are selected as a function of backup frequency, and backups are performed at intervals significantly shorter than the periods for which they are retained. Thus, future data protection offering faster backup and recovery times through the use of snapshots [1, 22], will have to be designed to handle significant data churn, or employ these mechanisms selectively.

We summarize the most important observations of our study in Table 1. Note that a *policy* (see Section 2.2) refers to a predefined set of configuration parameters specific to an application. The rest of the paper is organized as follows. In Section 2, we provide an overview of the evolution of backup systems. Section 3 describes the dataset used in this study. Sections 4 through 6 present our analysis results on backup domain configuration, job scheduling, and data growth, respectively. Finally, we discuss directions for research on next-generation data protection systems, supported by our findings, in Section 7, and conclude in Section 8.

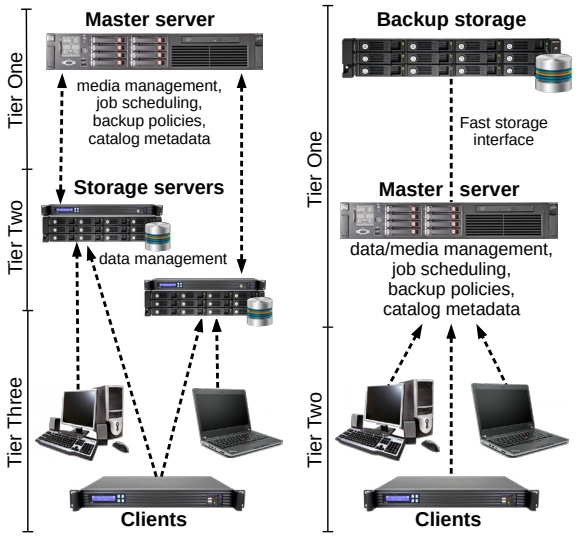
2 Background

Formally, *backup* is the process of making redundant copies of data, so that it can be retrieved if the original copy becomes unavailable. In the past 30 years, however, data growth coupled with capacity and bandwidth limitations have triggered a number of paradigm shifts in the way backup is performed. Recently, data growth trends have once again prompted efforts to rethink backup [1, 9, 20, 22, 27]. This section underlines the importance of field studies in this process (Section 2.1), putting our study in context, and describes the architecture of modern backup systems (Section 2.2).

2.1 Evolution of backup and field studies

In the early 1990s, backup consisted of using simple command-line tools to copy data to/from tape. A number of studies tested and outlined the shortcomings of these contemporary backup methods [38, 54, 69, 70]. The limitations of this approach, which included scaling, archive management, operating on online systems, and completion time, were subsequently addressed sufficiently by moving to a client-server backup model [8, 11, 15, 16]. In this model, job scheduling, policy configuration, and archive cataloging were all unified at the server side.

In the early 2000s, deduplicating storage systems were developed [53, 67], which removed data redundancy, lowering the cost of backup storage. Subsequently, Wallace et al. [66] published a study that aims to characterize backup storage characteristics by looking at the contents and workload of file systems that store images produced by backup applications such as NetBackup. A large body of work used their results to simulate deduplicating backup systems more realistically [41, 43, 44, 57, 62], and was built on the motivation provided by the study's results [40, 42, 46, 58]. The authors analyze weekly reports from appliances, while we analyze reports from the backup application, which has visibility within the archives and the jobs that created them. However, the two studies overlap in three points. First, the deduplication ratios reported for backups confirm our findings. Second, we report backup data retention as a configuration parameter, while they report on file age, two distri-



(a) 3-tier architecture (b) 2-tier architecture

Figure 1: Architecture of a modern backup domain.

butions that overlap for popular values. Third, the average job sizes we report are 5-8 times smaller than the file sizes reported in their study, likely because they take into account all files in the file system storing the backup images. Overlaps between our study and previous work are summarized in Table 1.

Recently, an ongoing effort has been initiated in the industry to redefine enterprise data protection as a response to modern data growth rates and shorter backup windows [12, 18, 65]. Proposed deviations from the traditional model rely on data snapshots, trading management complexity for faster job completion rates [22], and a paradigm shift from backup to data protection policies, in which users specify constraints on data availability as opposed to backup frequency and scheduling [1]. The latter paradigm allows the system to make decisions on individual policy parameters that can increase global efficiency, while keeping misconfigurations to a minimum. In this direction, previous work leverages predictive analytics to configure backup systems [9, 20, 25]. We believe that all this work is promising, and that a study characterizing the configuration and evolution of backup systems over time could aid in developing new approaches and predictive models that ensure backup systems meet their goals timely, while efficiently utilizing their resources.

2.2 Anatomy of modern backup systems

Modern *backup domains* typically consist of three tiers of operation: a master server, one or more storage servers, and several clients, as shown in Figure 1a. The domain's *master server* maintains information on backup images and backup policies. It is also responsible for scheduling and monitoring backup jobs, and assigning them to storage servers. *Storage servers* manage storage media, such as tapes and hard drives, used to archive backup images. By abstracting storage media management in this way, clients can send data directly to their corresponding storage server, avoiding a bandwidth bottleneck at the master server. Finally, domain *clients* can be desktops, servers, or virtual machines generating data that is protected by the backup system against failures. In an alternative 2-tiered architecture model (Figure 1b), the storage servers are absent and the storage media are directly managed by the master server. The majority of enterprise backup software today, including Symantec NetBackup, support the 3-tiered model [3, 5, 13, 17, 21, 28, 32, 60, 68].

Performing a backup generally consists of a sequence of operations, each of which is executed as an independent *job*. Such jobs include: *snapshots* of the state of data at a given point in time, copying data into a backup image as part of a *full backup*, copying modified data since the last backup as part of an *incremental backup*, restoring data from a backup image as part of a *recovery operation*, and managing backup images or backing up the domain's configuration as part of a *management operation*. These jobs are typically employed in a predefined order. For example, a full backup may be followed by a management operation that deletes backup images past their retention periods.

To be consistently backed up, or provide point-in-time recovery guarantees, business applications may require specific operations to take place. In these scenarios, backup products offer predefined *policies* that are specific to individual applications. For instance, a Microsoft Exchange Server policy will also backup the transaction log, to capture any updates since the backup was initiated. Users can further configure policies to specify the characteristics of backups jobs, such as their frequency and retention rate.

3 Dataset Information

Our analysis is based on *telemetry reports* collected from customer installations of a commercial backup product, Symantec NetBackup [61], in enterprise and regular production environments. Reports are only collected from customers who opted to participate in the telemetry program, so our dataset represents a fraction of the customer base. The reports contain no personal identifiable information, or details about the data being backed up.

Report types. Each report in our dataset belongs to exactly one of three types: installation, runtime, or domain report. Reports of different types are collected at distinct points in the lifetime of a backup domain. *Installation*

Report type	Metrics used in study
Installation	Installation time
Runtime report	Job information: starting time, type, size, number of files, client policy, deduplication ratio, retention period
Domain report	Number and type of policies, number of clients, number of storage media, number of storage servers and appliances

Table 2: Telemetry report metrics used in the study.

reports are generated when the backup software is successfully installed on a server, and can be used to determine the time each server of a domain first came online. *Runtime reports* are generated and transmitted on a weekly basis from online domains, and contain daily aggregate data about the backup jobs running on the system. *Domain reports* are also generated and transmitted on a weekly basis, and report daily aggregate metrics that describe the configuration of the backup domain. The telemetry report metrics used in this study are summarized in Table 2.

Dataset size. The telemetry reports in our dataset were collected over the span of 3 years (January 2012 to December 2014), across two major versions of the NetBackup software. We collected 1 million reports from over 40,000 server installations deployed in 124 countries, on most modern operating systems.

Monitoring duration. The backup domains included in our study were each monitored for 5.5 months on average, and up to 32 months. We elaborate on our strategy for excluding some of the domains from our analysis in Section 4.1. Note that the monitoring time is not always equivalent to the total lifetime of the domain, as many of these domains were still online at the time of this writing.

Architecture. While NetBackup supports the 3-tiered architecture model, only 35% of domains in our dataset use dedicated storage servers. The remaining domains omit that layer, opting for a 2-tier system instead. Additionally, while backup software can be installed on any server, storage companies also offer Purpose-Built Backup Appliances (PBBAs) [33]. 31% of domains in our dataset represent this market by deploying NetBackup on Symantec PBBAs.

4 Domain configuration

This section analyzes the way backup domains are configured with regards to their clients and backup policies. We use the periodic telemetry reports to quantify the growth rate of the number of clients and policies across domains, and characterize the diversity of policy types based on the type of data and applications they protect.

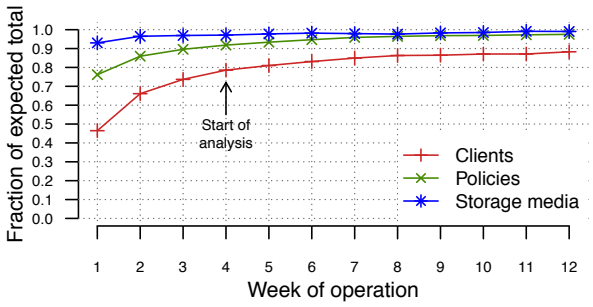


Figure 2: The average number of clients, policies, and storage media for a given week of operation, as a fraction of the expected total, i.e. the overall mean. We begin our analysis on the fourth week of operation, when these quantities become relatively stable.

4.1 Initial configuration period

Observation 1: *Backup domains take at least 3 weeks to reach a stable configuration after installation.*

The number of clients, policies, and storage media are three characteristic factors of a backup domain's configuration. These numbers fluctuate as resources are added to, or removed from the domain. As we monitor domains since their creation, we find the number of clients, policies, and storage media to be initially close to zero, and then increase rapidly until the domain is properly configured. After this initial configuration period, variability for these numbers tends to be low over the lifetime of each domain, with standard deviations less than 16% of the corresponding mean.

To avoid having the initial weeks of operation affect our results, we exclude them from our analysis. To estimate the average configuration period length, we analyze the number of clients, policies, and storage media in a backup domain as a fraction of the overall mean, i.e. the expected total. In Figure 2, we report the average fractions for all domains that have been monitored for more than 16 weeks. For example, a fraction of 0.47 for the number of clients during the first week of operation, implies that the number of clients at that time is 47% of the domain's expected total. With the exception of storage media, which seem to be added to backup domains from their first week of operation, we find that the number of clients and policies tends to be significantly lower for the first 3 weeks of operation. As a result, we choose to start our analysis from the fourth week of operation.

4.2 Client growth rate

Observation 2: *The number of clients in a domain increases by an average of 7 clients every 3.7 months.*

Clients are the producers of backup data, and the con-

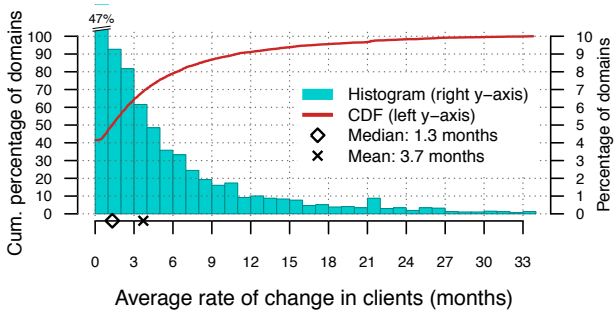


Figure 3: Distribution of the average rate at which the number of clients changes, across all domains in our dataset. On average, 93% of client population changes are attributed to the addition of clients.

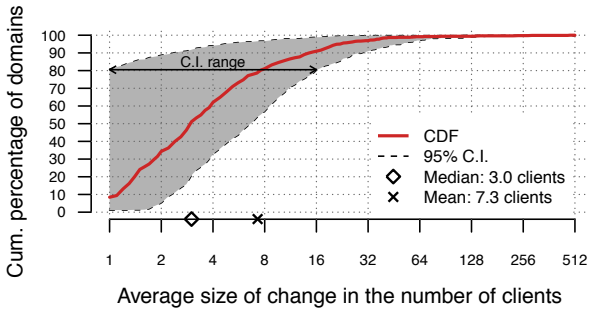


Figure 4: Distribution of the average number of clients involved in each change of a domain's client population, across all domains in our dataset. The 95% confidence intervals (C.I.) for each domain's average are also shown.

sumers of said data during recovery. As a result, the number of jobs running on a backup domain is directly proportional to the number of clients in the domain, deeming it important to quantify the rate at which their population grows over time.

Once the initial configuration period for a backup domain has elapsed, we find that clients tend to be added to, or removed from the domain in groups. Therefore, we characterize a domain's client population growth by quantifying the average rate of change in the client population, the sign indicating an increase or decrease in the population, and size of each change.

To estimate the rate at which the number of clients change, we extract inter-arrival times between changes through change-point analysis [37], a cost-benefit approach for detecting changes in time series. Then, we estimate the average rate of change for a domain as the average of these inter-arrival times. In Figure 3, we show the distribution of the average rates of change, i.e. the average number of months between changes in the number of clients across domains. For 42% of backup domains, the number of clients remains fixed after the first 3 weeks of operation, while on average the number of clients in a domain changes every 3.7 months. Overall, we find no strong correlation between the rate of change in the number of clients, and the domain's lifetime.

We further analyze the sign and size of each population change. Of all events in which a domain's client population changes, 93% are attributed to the addition of clients. However, 78% of domains never remove clients. Regarding the size of each change, Figure 4 shows the distribution of the average number of clients involved in each change, across all domains in our study. On average, a domain's population changes by 7.3 clients at a time. The average standard deviation of the number of clients over time is 13.1% of the corresponding expected value, indicating low variation overall. However, the 95% confidence intervals (C.I.) for each mean (Figure 4), suggest that growth spurts as large as 2.16 times the average value are possible, as this is the width of the average 95% confidence interval.

Policy category	Domains with at least 1 policy
File and block policy	61.24%
Database policy	20.34%
Virtual machine policy	15.13%
Application policy	13.52%
Metadata backup policy	**31.93%**

Table 3: Percentage of backup domains with at least one policy of a given category. Less than a third of domains protect the master server using a metadata backup policy.

4.3 Diversity of protected data

Observation 3: *82% of backup domains protect one type of data, and only 32% of domains effectively protect the master server's state and metadata.*

To provide consistent online backups, backup products offer optimizations for different application types, implemented as dedicated policy types [14, 23, 59]. For our analysis, we partitioned these policy types into four categories. *File and block policies* are specifically tailored for backing up raw device data blocks, or file and operating system data and metadata, e.g. from NTFS, AFS, or Windows volumes. *Database policies* are designed to provide consistent online backups for specific database management systems, such as DB2 and Oracle. *Virtual machine policies* are tuned to backup and restore VM images, from virtual environments such as VMware or Hyper-V. *Application policies* specialize in backing up state for client-server applications, such as Microsoft Exchange and Lotus Notes. Finally, a *metadata backup policy* can be setup to backup the master server's state.

In Table 3, we show the probability that at least one policy of a given category will be present in a backup domain. Since domains may deploy policies from multiple categories, these percentages add up to more than 100%.

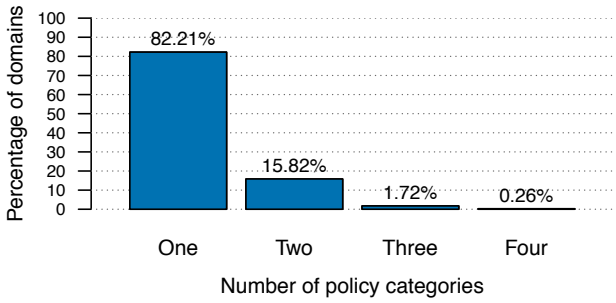


Figure 5: Distribution of the number of policy categories per backup domain. The metadata backup policy category is not accounted for in these numbers.

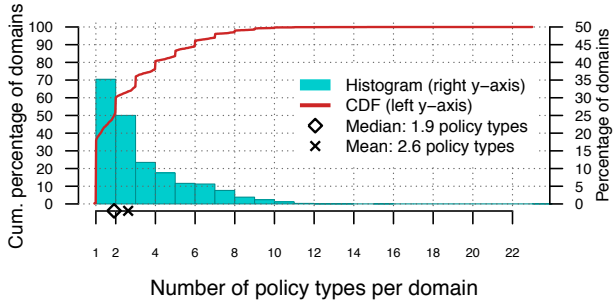


Figure 6: Distribution of the number of policy types per backup domain, across all domains in the study. More than 25 distinct NetBackup policy types are present in the telemetry data.

Surprisingly, we find that only 32% of backup domains register a metadata backup policy to protect the master server's data. While the remaining domains may employ a different mechanism to backup the master server, guaranteeing no data inconsistencies while doing so is challenging. In any case, this result suggests that automatically configured metadata backup policies should be a priority for future backup systems.

We also look into the number of policy categories represented by each domain's policies, to gauge the diversity in the types of protected data. Interestingly, Figure 5 shows that 82% of domains deploy policies of a single category (excluding metadata backup policies), and the remaining domains mostly use policies of two distinct categories. We further examine the number of distinct policy types that are deployed in each domain. As shown in Figure 6, domains tend to make use of a small number of policy types. Specifically, 61% of the domains deploy policies of only one, or two distinct types.

4.4 Backup policies

Observation 4: *After the initial configuration period, the number of policies in a domain remains mostly fixed and 79% of clients subscribe to a single policy each.*

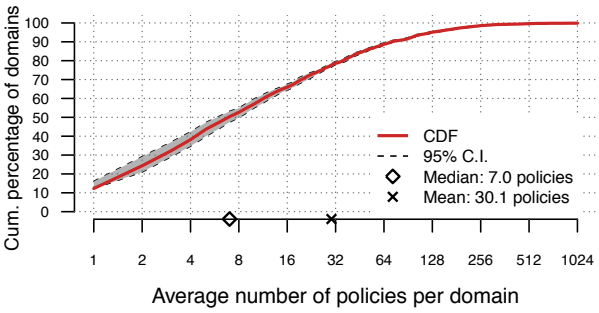


Figure 7: Distribution of the average number of policies per backup domain. The 95% confidence intervals for each average are also shown. Overall, the number of policies remains stable over the lifetime of a domain.

Following from Section 4.2, the policies in a backup domain, along with the number of clients, are indicative of the domain's load. Recall from Section 2.2, that clients subscribe to policies which determine the characteristics of backup jobs. Therefore, it is important to quantify both the number of policies in a domain and the characteristics of each, to effectively characterize the domain's workload. We defer an analysis of job characteristics to the remainder of the paper, and focus here on the number of policies in each domain.

In Figure 7, we show the distribution of the average number of policies in a given backup domain, across all domains in our dataset. Overall, we find that once the initial configuration period is complete, the number of backup policies in a domain remains mostly stable. Specifically, the expected width of the 95% confidence interval is 2.5% of the average number of policies.

Figure 7 also shows that the average backup domain carries 30 backup policies, while 5% of domains carry over 128. While each policy may represent a group of clients with specific data protection needs, we find that individual clients usually subscribe to a single policy. In Figure 8, we show the distribution of the average number of policies that each client subscribes to. More than 79% of clients belong to only one policy, while 16% spend some or most of their time unprotected (less than one policy on average). The latter result, coupled with the large number of policies in backup domains and the fact that clients are added to a domain in groups (Section 4.2), suggests that manual policy configuration might not be ideal as a domain's client population inflates over time.

5 Job scheduling

While the master server can reorder policy jobs to increase overall system efficiency, it adheres to user preferences that dictate when, and how often a job should be scheduled. This section looks into the way that these pa-

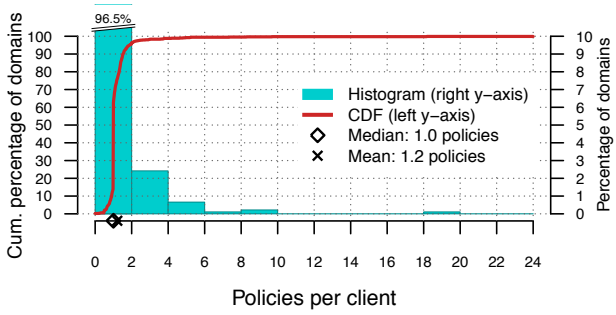


Figure 8: Distribution of the average number of policies that a domain client subscribes to. Overall, 79% of clients subscribe to one policy, while 16% spend some or most time unprotected by a policy ($x < 1$).

Job type	Percentage of jobs
Incremental Backups	45.27%
Full Backups	31.20%
Snapshot Operations	12.61%
Management Operations	10.12%
Recovery Operations	0.80%

Table 4: Breakdown of all jobs in the dataset by type.

rameters are configured by users across backup domains, and the workload generated in the domain as a result.

5.1 Job types

Recall from Section 2.2 that policies consist of a predefined series of operations, each carried out by a separate job. We collected data from 209.5 million jobs, and we group them in five distinct categories: full and incremental backups, snapshots, recovery, and management operations. In Table 4, we show a breakdown of all jobs in our dataset by job type. Across all monitored backup domains, we find that 76% of jobs perform data backups, having processed a total of 1.64 Exabytes of data, while 13% of jobs take snapshots of data. On the other hand, less than 1% of jobs are tasked with data recovery, having restored a total of 5.12 Petabytes of data. Finally, 10% of jobs are used to manage backup images, e.g. migrate, duplicate, or delete them. Due to the data transfer of backup images, these jobs processed 4.88 Exabytes of data. We analyze individual job sizes in Section 6.

5.2 Scheduling frequency

Observation 5: *Full backups tend to occur every 5 days or fewer. Recovery operations occur for few domains, on a weekly or monthly basis.*

A factor indicative of data churn in a backup domain is the rate at which jobs are scheduled to backup, restore, or manage backed-up data. To quantify the scheduling frequency of different job types for a given domain, we

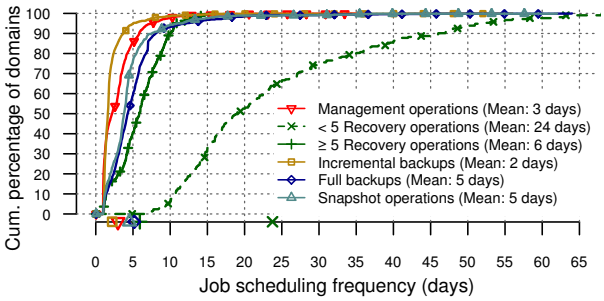


Figure 9: Distribution of the average scheduling frequency of different job types across backup domains. Recovery operations are broken into two groups of domains with more, and less than 5 recovery operations each. Despite being of similar size, the characteristics of each group differ significantly.

rely on the starting times of individual jobs. Specifically, starting times are used to estimate the average occurrence rate of different jobs of each domain policy, on individual clients. In Figure 9, we show the distributions of the scheduling frequency of different job types across backup domains.

Overall, we find that the average frequency of recovery operations differs depending on their number. In Figure 9, we show the distributions of the recovery frequency for two domain groups having recovered data more, and less than 5 times. The former group consists of 337 domains that recovered data 17 times on average, and the latter consists of 262 domains with 3 recovery operations on average. By definition, our analysis excludes an additional 676 domains that initiate recovery only once. For domains with multiple events, the distribution of their frequency spans 1-2 weeks, with an average of 6 days. On the other hand, domains with fewer recovery operations perform them significantly less frequently, up to 2 months apart and every 24 days on average. Since recovery operations are initiated manually by users, we have no accurate way of pinpointing their cause. These results, however, suggest that frequent recovery operations may be attributed to disaster recovery testing, while infrequent ones may be due to actual disasters. Interestingly, both domain groups are equally small, but when domains with a single recovery event are factored in, the group of infrequent recovery operations doubles in size.

In the case of backup jobs, the general belief is that systems in the field rely on weekly full backups, complemented by daily incremental backups [11, 36, 67]. Our results confirm this assumption for incremental backups, which take place every 1-2 days in 81% of domains. Daily incremental backups are also the default option in NetBackup. For full backups, however, our analysis shows that only 17% of domains perform them every 6-8

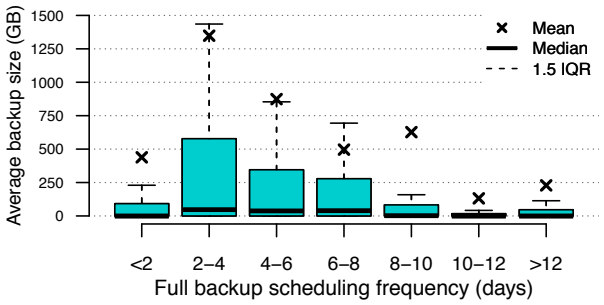


Figure 10: Tukey boxplots (without outliers) that represent the average size of full backup jobs, for different job scheduling frequencies. Means for each boxplot are also shown. Frequent full backups seem to be associated with larger job sizes, suggesting that they may be preferred as a response to high data churn.

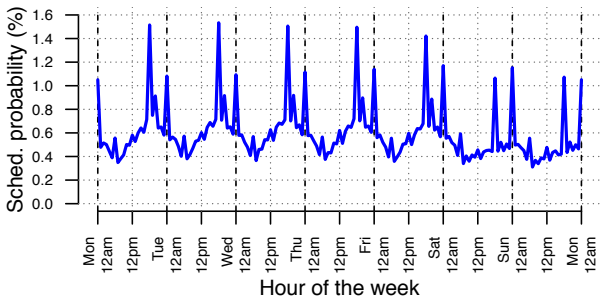


Figure 11: Probability density function for scheduling policy jobs at a given hour of a given day of the week. Policies tend to be configured using the default scheduling windows at 6pm and 12am, resulting in high system load during those hours.

days on average. Instead, the majority of domains perform full backups more often: 15% perform them every 1-2 days, and 57% perform them every 2-6 days. This is despite the fact that weekly full backups is the default option. As expected, management operations take place on a daily or weekly basis, since they usually follow (or precede) an incremental or full backup operation. Snapshot operations display a similar trend to full backups, as they are mostly used by clients in lieu of the latter.

Of the 65% of domain policies that perform full backups every 6 days or fewer, only 33% also perform incremental backups at all. On the other hand, 76% of policies that perform weekly full backups also rely on incremental backups. To determine whether full backups are performed frequently to accommodate high data churn, we group average full backup sizes per client policy according to their scheduling frequency, and present the results as a series of boxplots in Figure 10. Note that regardless of frequency, full backups tend to be small (medians in the order of a few gigabytes), due to the efficiency of deduplication. However, the larger percentiles of each distribution show that larger backup sizes tend to occur when full backups are taken more frequently than once per week. While this confirms our assumption of high data churn for a fraction of the clients, the remaining small backup sizes could also be attributed to overly conservative configurations, a sign that policy auto-configuration is an important feature for future data protection systems.

5.3 Scheduling windows

Observation 6: *Users prefer default scheduling windows during weekdays, resulting in nightly bursts of activity. Default values are overridden, however, to avoid scheduling jobs during the weekend.*

Another important factor for characterizing the workload of a backup system is the exact time jobs are scheduled. A popular belief is that backup operations take place late at night or during weekends, when client systems are expected to be idle [15, 66]. In Figure 11, we show our findings for all the jobs in our dataset. The presented density function was computed by normalizing the number of jobs that take place in a given domain, to prevent domains with more jobs from affecting the overall trend disproportionately. We note that this normalization had minimal effect on the result, which suggests that the presented trend is common across domains.

The hourly scheduling frequency is similar for each day, although there is less activity during the weekend. We also find that the probability of a job being scheduled is highest starting at 6pm and 12am on a weekday. We attribute the timing of job scheduling to customers using the default scheduling windows suggested by NetBackup, which start at 6pm and 12am every day. The choice to exclude weekends, however, seems to be an explicit choice of the user. This result suggests that automated job scheduling, where the only constraints would be to leverage device idleness [4, 26, 48], would be more practical, allowing the system to schedule jobs so that such activity bursts are avoided.

While Figure 11 merges all job types, different jobs exhibit different scheduling patterns, as shown in Figure 9. Our data, however, does not allow a matching of job types to scheduling times at a granularity finer than the day on which the job was scheduled. Thus, we partition jobs based on their type, and in Figure 12 we show the probability that a job of a given type will be scheduled on a given day of the week. We find that incremental backups are scheduled to complement full backups, as they tend to get scheduled from Monday to Thursday, while full backups are mostly scheduled on Fridays. Note that the latter does not contradict our previous result of full backups that take place more often than once a week,

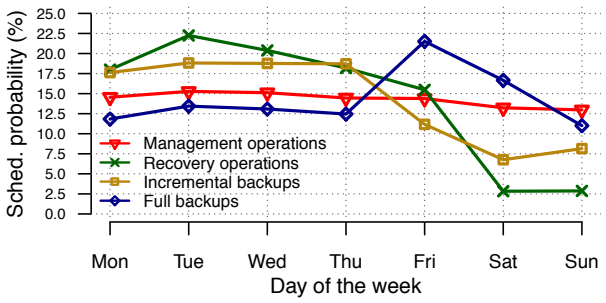


Figure 12: Probability of a policy job occurring on a given day of the week, based on its type. Incremental backups tend to be scheduled to complement full backups, while users initiate recovery operations more frequently at the beginning of the week.

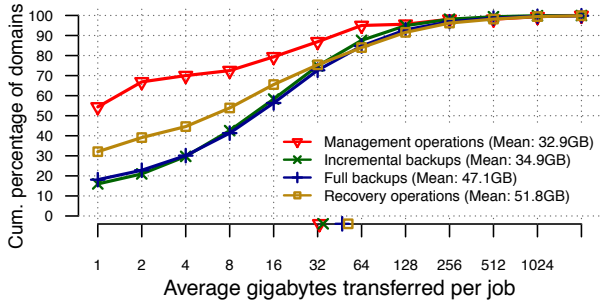


Figure 13: Distribution of the average job size of a given job type across backup domains, after the data has been deduplicated at the client side. Incremental backups resemble full backups in size.

as the probability of scheduling full backups any other day is still comparatively high. Recovery operations also take place within the week, with a slightly higher probability on Tuesdays (which we confirmed as not related to Patch Tuesday [49]). Finally management operations do not follow any particular trend and are equally likely to be scheduled on any day of the week.

6 Backup data growth

Characterizing backup data growth is crucial for estimating the amount of data that needs to be transferred and stored, which allows for efficient provisioning of storage capacity and bandwidth. Towards this goal, we analyze the sizes and number of files of different job types, and their deduplication ratios across backup domains. Finally, we look into the time that backup data is retained.

6.1 Job sizes and number of files

Observation 7: *Incremental and full backups tend to be similar in size and files transferred, due to the effectiveness of deduplication, or misconfigurations. Recovery jobs restore either a few files, or entire volumes.*

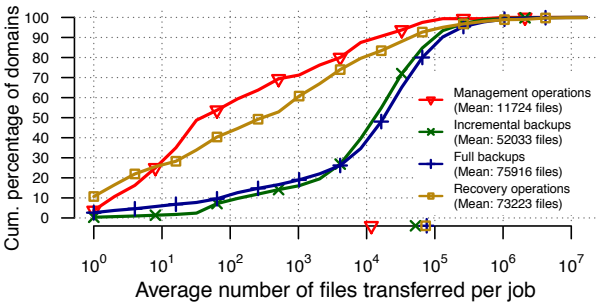


Figure 14: Distributions of average number of files transferred per job, across different job types. The trends are consistent with those for job sizes (Figure 13).

An obvious factor when estimating a domain's data growth is the size of backup jobs. In Figure 13, we show the distributions of the average number of bytes transferred for different job types across all domains, after the data has been deduplicated at the client. Averages for each operation are shown in the legend, and marked on the x axis. Snapshot operations are not included, as they do not incur data transfer.

Surprisingly, incremental backups resemble full backups in size. Although the distribution of full backups is skewed toward larger job sizes, 29% of full backups on domains that also perform incremental backups tend to be equal or smaller in size than the latter, 21% range from $1-1.5$ times the size of incremental backups, and the remainder range from $1.5-10^6$ times. We attribute the small size difference to three reasons. First, systems with low data churn can achieve high deduplication rates, which are common as we show in Section 6.2. Second, misconfigured policies or volumes that do not support incremental backups often fall back to full backups, as suggested by support tickets. Third, maintenance applications, such as anti-virus scanners, can update file metadata making unchanged files appear modified. Overall, the average backup job sizes in Figure 13 are 5-8 times smaller than the file sizes reported by Wallace et al. [66], likely due to their study considering the sizes of all files in the file system storing the backup images.

Since recovery operations can be triggered by users to recover an entire volume or individual files, the distribution of recovery job sizes is not surprising. 32% of recovery jobs restore less than 1GB, while the average job can be as large as 51GB. Finally, management operations, which consist mostly of metadata backups (95.7%), but also backup image (1.5%) and snapshot (2.8%) duplication operations, are much smaller than all other operations, as expected.

Figure 14 shows the distributions of the average number of files transferred for different job types in each domain. Similar to job sizes, the average number of files transferred per incremental backup is 31% smaller than

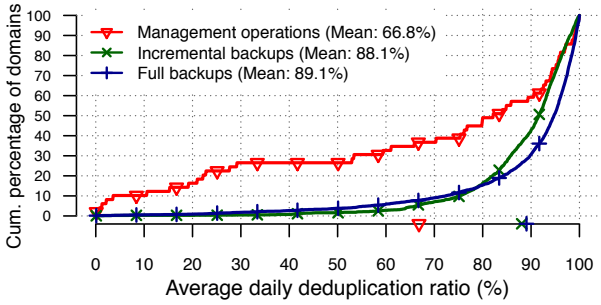


Figure 15: Distributions of the average daily deduplication ratio of different job types, across backup domains. Incremental and full backups observe high deduplication ratios, while the uniqueness of metadata backups (management operations) makes them harder to deduplicate.

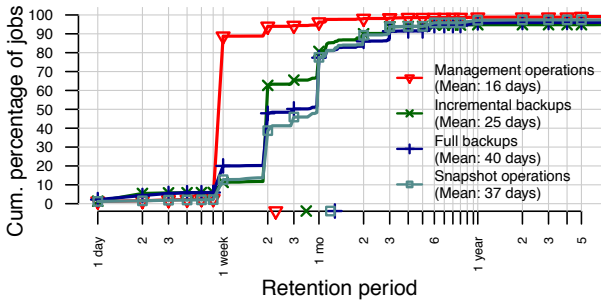


Figure 16: Distributions of retention period lengths for different job types. 3% of jobs have infinite retention periods. Incremental backups are typically retained for almost half the time of full backups, the majority of which are retained for months.

that for full backups, and both job types are characterized by similar CDF curves. Recovery operations transfer as many files as full backups on average, yet the majority transfer fewer than 200 files. This is in line with our results on recovery job sizes. Given that large recovery jobs also occur less frequently, these results suggest that most recovery operations are not triggered as a disaster response, but rather to recover data lost due to errors, or to test the recoverability of backup images. Management operations, being mostly metadata backups, transfer significantly fewer files than other job types on average.

6.2 Deduplication ratios

Observation 8: *Deduplication can result in the reduction of backup image sizes by more than 88%, despite average job sizes ranging in the tens of gigabytes.*

For clients that use NetBackup's deduplication solution, we analyzed the daily deduplication ratios of jobs, i.e. the percentage by which the number of bytes transferred was reduced due to deduplication. Figure 15 shows the distributions of the average daily deduplication ratio for management operations, full, and incremental backups across backup domains. Recovery and snapshot jobs are not included as the notion of deduplication does not apply. Since deduplication happens globally across backup images, deduplication ratios for backups tend to increase after the first few iterations of a policy. In general, sustained deduplication ratios as high as 99% are not unusual. Across all domains in our dataset, however, the average daily deduplication ratio is 88-89%, for both full and incremental backups. It is interesting to note that despite such high deduplication ratios, jobs in the order of tens of gigabytes are common (Figure 13), suggesting that even for daily incremental jobs, the actual job sizes are an order of magnitude larger in size. These results are in agreement with previous work [66], which reports average deduplication ratios of 91%.

Finally, for management operations the average deduplication ratio is 68%. Since only 1.1% of domains that use deduplication enable it for management operations, we do not attach much importance to this result. For the reported domains, however, it can be attributed to the uniqueness of metadata backups, which do not share files with other backup images on the same backup domain and consist of large binary files.

6.3 Data retention

Observation 9: *Incremental backups are retained for weeks, while full backups are retained for months and retention depends on their scheduling frequency.*

Another factor characteristic of backup storage growth is the retention time for backup images, which is a configurable policy parameter. Once a backup image expires, the master server deletes it from backup storage. We have analyzed the retention periods assigned to each job in our telemetry reports, and show the distributions of retention period lengths for different job types in Figure 16. Our initial observation is that job retention periods coincide with the values available by default in NetBackup, although users can specify custom periods. These values range from 1 week to 1 year, and correspond to the steps in the CDF shown. While federal laws, such as HIPAA [63] and FoIA [64], require minimum retention from a few years up to infinity for certain types of data. In our case, 3% of jobs are either assigned custom retention periods longer than 1 year, or are retained indefinitely. On the other extreme, only 3% of jobs are assigned custom retention periods shorter than 1 week. Previous work confirms our findings, by reporting similar ages for backup image files [66].

In particular, management operations (metadata backups and backup image duplicates) are mostly retained for 1 week. Incremental backups are mostly retained for 2 weeks, the default option. Full backups and snapshots,

on the other hand, are more likely retained for months. Overall, 94% of jobs select a preset retention period from NetBackup's list, and 35% of jobs keep the default suggestion of 2 weeks. This suggests that the actual retention period length is not a crucial policy parameter.

Finally, we find a strong correlation (Pearson's $r = 0.53$) between the length of retention periods for full backups, and the frequency with which they take place. Specifically, we find that clients taking full backups less frequently retain them for longer periods of time. On the other hand, no such correlation exists for management operations and incremental backups. This is because almost all data resulting from a management operation is retained for 1 week (Figure 16), and almost all incremental backups are performed with a frequency of 1-2 days apart (Figure 9). The correlation of retention period length and frequency of full backup operations, coupled with the preference for default values, may suggest that retention periods are selected as a function of storage capacity, or that they are at least limited by that factor.

7 Insight: next-generation data protection

This section outlines five major directions for future work on data protection systems. In each case, we identify existing literature and describe how our findings encourage future work.

Automated configuration and self-healing. To alleviate performance and availability problems of data protection systems, existing work uses historical data to perform automated storage capacity planning [9], data prefetching and network scheduling [25]. Our findings support this line of work. We have shown that backup domains grow in bursts, and client policies are either configured using default values, misconfigured, or not configured at all. As a result, clients are left unprotected, jobs are scheduled in bursts, and users are not warned of imminent problems. To enable automated policy configuration and self-healing data protection systems, further research is necessary.

Deduplication. Our findings confirm the efficiency of deduplication at reducing backup image sizes. We further show that in many systems, incremental backups are replaced by frequent full, deduplicated backups. This is likely due to the adoption of deduplication, which improves on incremental backups by looking for duplicates across all backup data in the domain. To completely replace incremental backups, however, it is necessary to improve on the time required to restore the original data from deduplicated storage, which directly affects recovery times. Currently, this is an area of active research [24, 35, 43, 50].

Efficient storage utilization. Our analysis shows that job retention periods are selected as a function of backup frequency, likely to ensure sufficient backup storage space will be available. Additionally, 31% of domains in our dataset use dedicated backup appliances (PBBAs), a market currently experiencing growth [33]. We believe that storage capacity in these dedicated systems should be utilized fully, and retention periods should be dynamically adjusted to fill it, providing the ability to recover older versions of data. In this direction, related work on stream-processing systems [29] could be adapted to the needs of backup data.

Accident insurance. Most recovery operations in our dataset appear to be small in both the number of files and bytes they recover, compared to their respective backups. This result suggests that recovery operations are mostly triggered to restore a few files, or to test the integrity of backup images. This motivates us to re-examine the requirement of instant recovery for backup systems as a problem of determining which data is more likely to be recovered, and storing it closer to clients [40, 45].

Content-aware backups. Data protection strategies can generate data at a rate up to 5 times higher than production data growth [1]. This is due to the practice of creating multiple copies and backing up temporary files used for test-and-development or data analytics processes, such as the Shuffle stage of MapReduce tasks [10]. Depending on the storage interface used, it might be more efficient to recompute these datasets rather than restoring them from backup storage. Another challenge for contemporary backup software is detecting data changes since the last backup among PBs of data and billions of files [31]. By augmenting data protection systems to account for data types and modification events, we can potentially reduce the time needed to complete backup and restore operations.

8 Conclusion

We investigated an extensive dataset representing a diverse population of enterprise data protection systems to demonstrate how these systems are configured and evolved over time. Among other results, our analysis showed that these systems are usually configured to protect one type of data, and while their client population growth is steady and bursty, their backup policies don't change. With regards to job scheduling, we find that the popularity of default values can have an adverse effect on the efficiency of the system by creating bursty workloads. Finally, we showed that full and incremental backups tend to be similar in size and number of files, as a result of efficient deduplication and misconfigurations. We hope that our data and the proposed areas of future research will enable researchers to simulate realistic scenarios for building next generation data protection systems that are easy to configure and manage.

Acknowledgments

The study would not be possible without the telemetry data collected by Symantec's NetBackup team, and we thank Liam McNerney and Aaron Christensen for their invaluable assistance in understanding the data. We also thank the four anonymous reviewers and our shepherd, Fred Douglis, for helping us improve our paper significantly. Finally, we would like to thank Petros Efstathopoulos, Fanglu Guo, Vish Janakiraman, Ashwin Kayyoor, CW Hobbs, Bruce Montague, Sanjay Sahwney, and all other members of Symantec's Research Labs for their feedback during the earlier stages of our study.

References

[1] ACTIFIO. Actifio Copy Data Virtualization: How It Works, August 2014.

[2] AGRAWAL, N., BOLOSKY, W. J., DOUCEUR, J. R., AND LORCH, J. R. A five-year study of file-system metadata. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies* (2007).

[3] ARCSERVE. arcserve Unified Data Protection. `http://www.arcserve.com`, May 2014.

[4] BACHMAT, E., AND SCHINDLER, J. Analysis of methods for scheduling low priority disk drive tasks. In *Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and modeling of computer systems* (2002).

[5] BACULA SYSTEMS. Bacula 7.0.5. `http://www.bacula.org`, July 2014.

[6] BAKER, M., HARTMAN, J. H., KUPFER, M. D., SHIRRIFF, K., AND OUSTERHOUT, J. K. Measurements of a Distributed File System. In *Proceedings of the 13th ACM Symposium on Operating Systems Principles* (1991).

[7] BENNETT, J. M., BAUER, M. A., AND KINCHLEA, D. Characteristics of Files in NFS Environments. In *Proceedings of the 1991 ACM SIGSMALL/PC Symposium on Small Systems* (1991).

[8] BHATTACHARYA, S., MOHAN, C., BRANNON, K. W., NARANG, I., HSIAO, H.-I., AND SUBRAMANIAN, M. Coordinating Backup/Recovery and Data Consistency Between Database and File Systems. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data* (2002), SIGMOD.

[9] CHAMNESS, M. Capacity Forecasting in a Backup Storage Environment. In *Proceedings of the 25th International Conference on Large Installation System Administration* (2011).

[10] CHEN, Y., ALSPAUGH, S., AND KATZ, R. Interactive Analytical Processing in Big Data Systems: A Cross-industry Study of MapReduce Workloads. *Proc. VLDB Endow. 5*, 12 (Aug. 2012), 1802–1813.

[11] CHERVENAK, A. L., VELLANKI, V., AND KURMAS, Z. Protecting File Systems: A Survey Of Backup Techniques. In *Proceedings of the Joint NASA and IEEE Mass Storage Conference* (1998).

[12] COMMVAULT SYSTEMS. Get Smart About Big Data: Integrated Backup, Archive & Reporting to Solve Big Data Management Problems, July 2013.

[13] COMMVAULT SYSTEMS INC. CommVault Simpana 10. `http://www.commvault.com/simpana-software`, April 2014.

[14] COMMVAULT SYSTEMS INC. CommVault Simpana: Solutions for Protecting and Managing Business Applications. `http://www.commvault.com/solutions/enterprise-applications`, April 2015.

[15] DA SILVA, J., GUDMUNDSSON, O., AND MOSSÉ, D. Performance of a Parallel Network Backup Manager, 1992.

[16] DA SILVA, J., AND GUTHMUNDSSON, O. The Amanda Network Backup Manager. In *Proceedings of the 7th USENIX Conference on System Administration* (1993), LISA.

[17] DELL INC. Dell NetVault 10.0. `http://software.dell.com/products/netvault-backup`, May 2014.

[18] DIMENSIONAL RESEARCH. The state of IT recovery for SMBs. `http://axcient.com/state-of-it-recovery-for-smbs`, Oct. 2014.

[19] DOUCEUR, J. R., AND BOLOSKY, W. J. A Large-scale Study of File-system Contents. In *Proceedings of the 1999 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (1999).

[20] DOUGLIS, F., BHARDWAJ, D., QIAN, H., AND SHILANE, P. Content-aware Load Balancing for Distributed Backup. In *Proceedings of the 25th International Conference on Large Installation System Administration* (2011), LISA.

[21] EMC CORPORATION. EMC NetWorker 8.2. `http://www.emc.com/data-protection/networker.htm`, July 2014.

[22] EMC CORPORATION. EMC ProtectPoint: Protection Software Enabling Direct Backup from Primary Storage to Protection Storage, 2014.

[23] EMC CORPORATION. EMC NetWorker Application Modules Data Sheet. `http://www.emc.com/collateral/software/data-sheet/h2479-networker-app-modules-ds.pdf`, January 2015.

[24] FU, M., FENG, D., HUA, Y., HE, X., CHEN, Z., XIA, W., HUANG, F., AND LIU, Q. Accelerating Restore and Garbage Collection in Deduplication-based Backup Systems via Exploiting Historical Information. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference* (2014).

[25] GIAT, A., PELLEG, D., RAICHSTEIN, E., AND RONEN, A. Using Machine Learning Techniques to Enhance the Performance of an Automatic Backup and Recovery System. In *Proceedings of the 3rd Annual Haifa Experimental Systems Conference* (2010), SYSTOR.

[26] GOLDING, R., BOSCH, P., STAELIN, C., SULLIVAN, T., AND WILKES, J. Idleness is not sloth. In *Proceedings of the USENIX 1995 Technical Conference Proceedings* (1995), TCON'95.

[27] HEWLETT-PACKARD. Rethinking backup and recovery in the modern data center, November 2013.

[28] HEWLETT-PACKARD COMPANY. HP Data Protector 9.0.1. `http://www.autonomy.com/products/data-protector`, August 2014.

[29] HILDRUM, K., DOUGLIS, F., WOLF, J. L., YU, P. S., FLEISCHER, L., AND KATTA, A. Storage Optimization for Large-scale Distributed Stream-processing Systems. *Trans. Storage 3*, 4 (Feb. 2008), 5:1–5:28.

[30] HSU, W. W., AND SMITH, A. J. Characteristics of I/O Traffic in Personal Computer and Server Workloads. Tech. rep., EECS Department, University of California, Berkeley, 2002.

[31] HUGHES, D., AND FARROW, R. Backup Strategies for Molecular Dynamics: An Interview with Doug Hughes. *Proc. USENIX ;login: 36*, 2 (Apr. 2011), 25–28.

[32] IBM CORPORATION. IBM Tivoli Storage Manager 7.1. `http://www.ibm.com/software/products/en/tivostormana`, November 2013.

[33] INTERNATIONAL DATA CORPORATION. Worldwide Purpose-Built Backup Appliance (PBBA) Market Revenue Increases 11.2% in the Third Quarter of 2014, According to IDC. `http://www.idc.com/getdoc.jsp?containerId=prUS25351414`, December 2014.

[34] IRON MOUNTAIN. Data Backup and Recovery Benchmark Report. `http://www.ironmountain.com/Knowledge-Center/Reference-Library/View-by-Document-Type/White-Papers-Briefs/I/Iron-Mountain-Data-Backup-and-Recovery-Benchmark-Report.aspx`, 2013.

[35] KACZMARCZYK, M., BARCZYNSKI, M., KILIAN, W., AND DUBNICKI, C. Reducing Impact of Data Fragmentation Caused by In-line Deduplication. In *Proceedings of the 5th Annual International Systems and Storage Conference* (2012).

[36] KEETON, K., SANTOS, C., BEYER, D., CHASE, J., AND WILKES, J. Designing for Disasters. In *Proceedings of the 3rd USENIX Conference on File and Storage Technologies* (2004), FAST.

[37] KILLICK, R., AND ECKLEY, I. A. changepoint: An R package for Changepoint Analysis. In *Journal of Statistical Software* (May 2013).

[38] KOLSTAD, R. A Next Step in Backup and Restore Technology. In *Proceedings of the 5th USENIX Conference on System Administration* (1991), LISA.

[39] LEUNG, A. W., PASUPATHY, S., GOODSON, G., AND MILLER, E. L. Measurement and Analysis of Large-scale Network File System Workloads. In *Proceedings of the USENIX 2008 Annual Technical Conference* (2008).

[40] LI, C., SHILANE, P., DOUGLIS, F., SHIM, H., SMALDONE, S., AND WALLACE, G. Nitro: A Capacity-Optimized SSD Cache for Primary Storage. In *Proceedings of the 2014 USENIX Annual Technical Conference* (2014), ATC.

[41] LI, M., QIN, C., LEE, P. P. C., AND LI, J. Convergent Dispersal: Toward Storage-Efficient Security in a Cloud-of-Clouds. In *Proceedings of the 6th USENIX Workshop on Hot Topics in Storage and File Systems* (2014), HotStorage.

[42] LI, Z., GREENAN, K. M., LEUNG, A. W., AND ZADOK, E. Power Consumption in Enterprise-scale Backup Storage Systems. In *Proceedings of the 10th USENIX Conference on File and Storage Technologies* (2012), FAST.

[43] LILLIBRIDGE, M., ESHGHI, K., AND BHAGWAT, D. Improving Restore Speed for Backup Systems that Use Inline Chunk-Based Deduplication. In *Proceeings of the 11th USENIX Conference on File and Storage Technologies* (2013), FAST.

[44] LIN, X., LU, G., DOUGLIS, F., SHILANE, P., AND WALLACE, G. Migratory Compression: Coarse-grained Data Reordering to Improve Compressibility. In *Proceedings of the 12th USENIX Conference on File and Storage Technologies* (2014), FAST.

[45] LIU, J., CHAI, Y., QIN, X., AND XIAO, Y. PLC-cache: Endurable SSD cache for deduplication-based primary storage. In *Mass Storage Systems and Technologies (MSST), 2014 30th Symposium on* (2014).

[46] MEISTER, D., BRINKMANN, A., AND SÜSS, T. File Recipe Compression in Data Deduplication Systems. In *Proceedings of the 11th USENIX Conference on File and Storage Technologies* (2013), FAST.

[47] MEYER, D. T., AND BOLOSKY, W. J. A Study of Practical Deduplication. In *Proceedings of the 9th USENIX Conference on File and Stroage Technologies* (2011).

[48] MI, N., RISKA, A., LI, X., SMIRNI, E., AND RIEDEL, E. Restrained utilization of idleness for transparent scheduling of background tasks. In *Proceedings of the 11th International Joint conference on Measurement and modeling of computer systems* (2009), SIGMETRICS.

[49] MICROSOFT CORPORATION. Understanding Windows automatic updating. `http://windows.microsoft.com/en-us/windows/understanding-windows-automatic-updating`.

[50] NG, C.-H., AND LEE, P. P. C. RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups. In *Proceedings of the 4th Asia-Pacific Workshop on Systems* (2013).

[51] OUSTERHOUT, J. K., DA COSTA, H., HARRISON, D., KUNZE, J. A., KUPFER, M., AND THOMPSON, J. G. A Trace-driven Analysis of the UNIX 4.2 BSD File System. In *Proceedings of the 10th ACM Symposium on Operating Systems Principles* (1985).

[52] PARK, N., AND LILJA, D. J. Characterizing Datasets for Data Deduplication in Backup Applications. In *Proceedings of the IEEE International Symposium on Workload Characterization* (2010), IISWC.

[53] QUINLAN, S., AND DORWARD, S. Venti: A New Approach to Archival Data Storage. In *Proceedings of the 1st USENIX Conference on File and Storage Technologies* (2002), FAST.

[54] ROMIG, S. M. Backup at Ohio State, Take 2. In *Proceedings of the 4th USENIX Conference on System Administration* (1990), LISA.

[55] ROSELLI, D., LORCH, J. R., AND ANDERSON, T. E. A Comparison of File System Workloads. In *Proceedings of the USENIX Annual Technical Conference* (2000).

[56] SATYANARAYANAN, M. A Study of File Sizes and Functional Lifetimes. In *Proceedings of the 8th ACM Symposium on Operating Systems Principles* (1981).

[57] SHIM, H., SHILANE, P., AND HSU, W. Characterization of Incremental Data Changes for Efficient Data Protection. In *Proceedings of the 2013 USENIX Annual Technical Conference* (2013), ATC.

[58] SMALDONE, S., WALLACE, G., AND HSU, W. Efficiently Storing Virtual Machine Backups. In *Proceedings of the 5th USENIX Workshop on Hot Topics in Storage and File Systems* (2013), HotStorage.

[59] SYMANTEC CORPORATION. Symantec NetBackup 7.6 Data Sheet: Data Protection. `http://www.symantec.com/content/en/us/enterprise/fact_sheets/b-netbackup-ds-21324986.pdf`, January 2014.

[60] SYMANTEC CORPORATION. Symantec NetBackup 7.6. `http://www.symantec.com/backup-software`, March 2015.

[61] SYMANTEC CORPORATION. Symantec NetBackup 7.6.1 Getting Started Guide. `https://support.symantec.com/en_US/article.DOC7941.html`, February 2015.

[62] TARASOV, V., MUDRANKIT, A., BUIK, W., SHILANE, P., KUENNING, G., AND ZADOK, E. Generating Realistic Datasets for Deduplication Analysis. In *Proceedings of the 2012 USENIX Annual Technical Conference* (2012), ATC.

[63] U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. The Health Insurance Portability and Accountability Act. `http://www.hhs.gov/ocr/privacy`.

[64] U.S. DEPARTMENT OF JUSTICE. The Freedom of Information Act. `http://www.foia.gov`.

[65] VANSON BOURNE. Virtualization Data Protection Report 2013 – SMB edition. `http://www.dabcc.com/documentlibrary/file/virtualization-data-protection-report-smb-2013.pdf`, 2013.

[66] WALLACE, G., DOUGLIS, F., QIAN, H., SHILANE, P., SMALDONE, S., CHAMNESS, M., AND HSU, W. Characteristics of Backup Workloads in Production Systems. In *Proceedings of the 10th USENIX Conference on File and Storage Technologies* (2012).

[67] ZHU, B., LI, K., AND PATTERSON, H. Avoiding the Disk Bottleneck in the Data Domain Deduplication File System. In *Proceedings of the 6th USENIX Conference on File and Storage Technologies* (2008).

[68] ZMANDA INC. Amanda 3.3.6. `http://amanda.zmanda.com`, July 2014.

[69] ZWICKY, E. D. Torture-testing Backup and Archive Programs: Things You Ought to Know But Probably Would Rather Not. In *Proceedings of the 5th USENIX Conference on System Administration* (1991), LISA.

[70] ZWICKY, E. D. Further Torture: More Testing of Backup and Archive Programs. In *Proceedings of the 17th USENIX Conference on System Administration* (2003), LISA.