Symantec Research Labs
RIPS 2014 Project Description

1. Introduction

Data is growing at a colossal rate with 2.5 Exa Bytes of data generated every day (2012 study). In addition to financial transactions, online shopping and banking, using smart devices, the popularity of social networks, etc., data protection strategies such as backup and restore also contribute to this data explosion. Data protection landscape has gone through a significant transformation in the last 5 years. Today, there is a multitude of technologies in this area such as tape-based backup, disk-based backup, continuous data protection, wide area replication, asynchronous/asynchronous mirroring, virtual tape libraries, data de-duplication, and data encryption. These are managed by system administrators to ensure business continuity. System administrators have to manage labor intensive and complex tasks including configuring these systems for optimal performance, resource allocation, and failure detection. To make the admin's job easy, recent trends have suggested automating some of these tasks.

For this project, we will focus on the task of predicting the "best" amount of storage capacity for backup systems, and based on this, predicting the time of when the backup will complete. Allocating more capacity than necessary can be cost inefficient, while not assigning enough can prove to be inefficient or detrimental to the business. For instance, if the backup workload is expected to generate utmost 1GB of data per month, a backup system with 100 GB spare capacity adds up unnecessary costs. On the other hand, if the workload generated 25GB a month, the same system would require additional storage in 4 months. In real systems however, the backup workload does not generate the same amount of data every month, and there are several factors that play a role in determining it.

What makes automated capacity prediction a significant contribution is that it can enable predicting when the backup will complete, which in turn can predict if recovery point objective (RPO) can be met. RPO is an important service level agreement (SLA) that always has to be met to ensure no data is lost in case of failures.

2. Background

There are a number of factors that play a role in capacity planning, and performance. In addition to the data that needs to be backed up, metadata information is also stored for bookkeeping. It provides backup history such as when a backup was initiated and completed, what the size of the backup was, and also helps identify where the backup image is stored, so that the latest image can be used for data recovery.

Bookkeeping information is also necessary for deduplication, a compression technique to eliminate redundancy. Briefly, deduplication works as follows to save capacity by eliminating redundancies in data. A fingerprint is stored for every extent of data. New data that needs to be stored is compared with these fingerprints. If a match is found, data is discarded. If no match is found, an entry is made for this extent in the fingerprint index. Deduplication is useful because it not only reduces the amount of storage required, but also reduces the bytes to be transferred over the network and subsequently the load on network bandwidth.

Performance depends on various factors, such as the number of clients generating the backup workload, the size of the backup workload, the number of devices, streams, and the network bandwidth.

3. Problem definition
   The goal of the project is to answer the following questions.
   a. Given past history of the size of the data backed up, when is my backup data expected to exceed the assigned storage capacity? In other words, when should I add more storage to the system, and how much?
   b. How do I predict this when there is no/not enough history? Or, how do I estimate how much storage I will need to begin with?
   c. Based on history, and my prediction of backup storage, when will my backup not meet my RPO SLA?
   d. How does deduplication affect this?

4. Expected outcome
   Task 1: Formalize a list of all inputs, constraints and outputs for the model.
   Task 2: Design a model to predict the size of the next backup, and time taken to complete the backup.
   Task 3: Evaluate the accuracy of the model with real-data/prototype
   Deliverable 1: A technical report detailing the model, experimental results and conclusions.
   Deliverable 2: Presentation to Symantec Research Labs on how to use the model, and lessons learnt.
   Deliverable 3: Any prototype source/software/tools written for this work.

5. Recommended reading
   1. Paper on capacity planning: Characteristics of Backup Workloads in Production workloads (http://www.research.rutgers.edu/~smaldone/pubs/backup-fast12.pdf).
   2. Sample Backup SLA: http://dedicatedserver.com/solutions/sla/products/data_center_backup.cfm
   3. SLA and disaster recovery: http://articles.techrepublic.com.com/5100-10878_11-6048880.html
   4. RTO and RPO: http://www.wikibon.org/Recovery_point_objective_/_recovery_time_objective_strategy
   5. Business continuity metrics: How much can you afford to lose? http://www.computerworld.com/printthis/2004/0,4814,92865,00.html