the Availability Digest

www.availabilitydigest.com

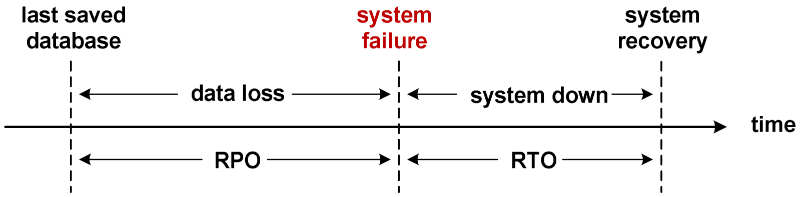
The Cost of RPO and RTO

September 2012

The purpose of availability analysis is to determine how to limit downtime and data loss. Both cost a corporation money and reputation. But improving them also costs money. Improvement generally means adding redundancy to the corporate systems. How does one balance the cost of availability improvement against the savings of reduced downtime and lost data?

For every application, the company should set certain objectives for lost downtime and lost data. The objective for lost downtime is called the Recovery Time Objective, or RTO. The objective for lost data is called the Recovery Point Objective, or RPO.

The techniques for minimizing downtime and lost data are, in general, largely independent. Redundancy is used to minimize downtime. Data replication is use to minimize lost data. Downtime is minimized by providing geographically dispersed redundant servers and storage. The faster the failover, the lower is RTO. Lost data is minimized by maintaining a copy of the data at a safe site. The shorter the replication latency (that is, the time from when a change is made to the production database to the time that it is made to the backup database), the lower is RPO.



Definition of RPO, RTO

In this article, we look at relationships between costs and savings; and we generate some rules of thumb for arriving at the best compromise. The analysis is the same for both RPO and RTO. Therefore, we focus on RPO as an example.

Recovery Point Objective (RPO)

RPO Architectures

Data can be protected against loss via many techniques, each with their own RPO and cost characteristics:

- The classic method is magnetic tape. All data from the last full or incremental backup is lost. RPO can be measured in days.

- Virtual tape writes magnetic tape images to disk instead of tape. This is much faster, so backups can be taken more frequently. RPO is typically measured in hours.
- Data changes can be replicated to a backup SAN by the SAN system. RPO is generally measured in minutes.
- Asynchronous replication replicates changes after they are made to the production database to a backup database. RPO is typically in the order of seconds.
- Synchronous replication ensures that a change is made on the backup database before the production database change completes. RPO is zero in this case.

The RPO/Cost Relationship

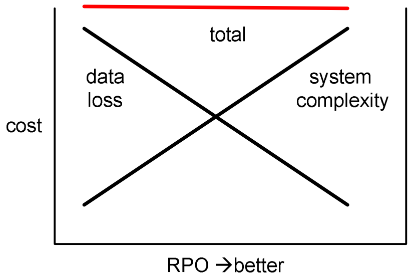
The cost of lost data goes down as RPO approaches zero. The cost of the system goes up as RPO approaches zero. By plotting these two cost curves and adding them, the total cost for any RPO solution can be estimated. The task is then to determine the least-cost solution. This will define the ideal RPO from a cost basis.

Let us consider some cases of RPO costs.

Example 1: Linear Costs

<u>Identical Linear Costs</u>

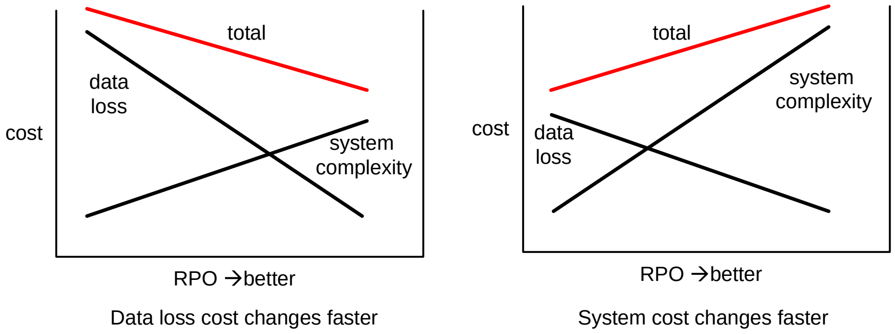
If the system cost increases linearly with RPO and the cost of lost data decreases with RPO at the same rate, the total cost is constant. It does not matter which solution is adopted from a cost viewpoint. In this case, choose the solution that minimizes RPO.



Example 1: Identical Linear Costs

<u>Different Linear Costs</u>

It the costs are linear but change at different rates, the cost that changes fastest wins. If the cost of lost data decreases at a faster rate than the system cost escalates, choose the minimum RPO. If the cost of the system increases faster than the cost of lost data decreases, choose the maximum RPO.



Data loss cost changes faster

System cost changes faster

Example 2: Different Linear Costs

Example 2: Exponential costs

We now look at a more difficult case, but one that is more representative of the real world. The costs are not linear, but always change in the same direction. The cost of lost data always decreases as RPO approaches zero, and the cost of the system always increases as RPO approaches zero. We get into some heavier mathematics here, but the conclusions are meaningful.

We assume that the minimum cost is somewhere between zero RPO and maximum RPO. Then the minimum cost is the point at which the slope of the cost curve is zero. This means that the derivative of the total cost with respect to RPO is zero (remember your calculus?):

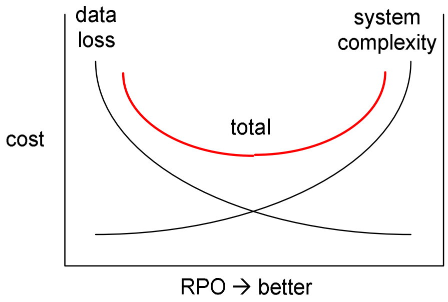
$$d(\text{total cost})/dRPO = d(\text{data cost}+\text{system cost})/dRPO = 0$$
$$d(\text{data cost})/dRPO = -d(\text{system cost})/dRPO$$

The slope of the data cost curve is positive (the cost of lost data increases with increasing RPO). The slope of the system cost curve is negative (the cost of the system decreases with increasing RPO). Therefore, the minimum cost point is that RPO at which the slopes of the data cost and system cost are equal.

Rule 1: *The minimum cost point is that RPO at which the slopes of the data cost and the system cost are equal.*

Exponential functions with cost of data loss increasing more rapidly than RPO increases

Let us now take the case of exponentially changing data loss cost and system cost in which the cost of lost data increases more rapidly than RPO increases. The cost curves can be represented as



$$\text{Cost of lost data} = d(e^{aRPO} - 1)$$
$$\text{Cost of system} \;= s(e^{-bRPO})$$

Then, the total cost is

$$\text{Total cost} = C = d(e^{aRPO} - 1) + s(e^{-bRPO})$$

We take the derivative of the total cost and set it to zero to find the RPO corresponding to zero slope:

$$dC/dRPO = ad(e^{aRPO}) – bs(e^{-bRPO}) = 0$$

$$e^{aRPO}/ e^{-bRPO} = bs/ad$$

$$e^{(b+a)RPO} = bs/ad$$

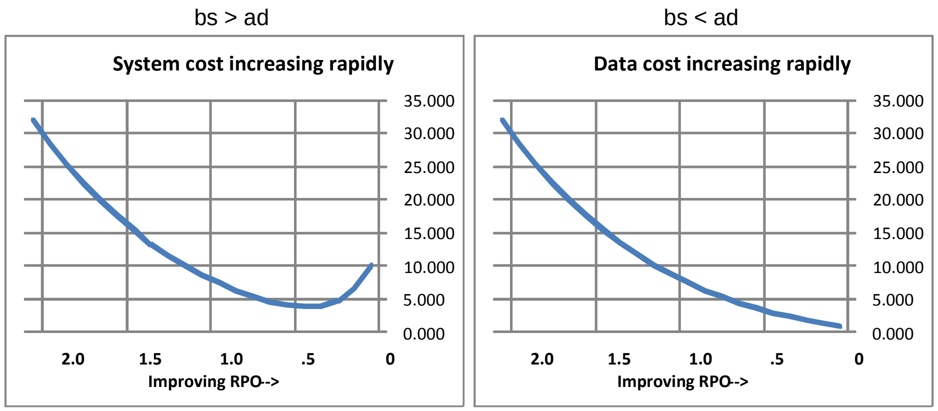
Taking the natural logarithm of both sides

$$(b+a)RPO = \ln(bs/ad)$$

Thus:

$$\text{minimum RPO} = \ln(bs/ad)/(b+a)$$

This result has an interesting interpretation. Note that *bs* must be greater than *ad.* Otherwise, *ln(bs/ad)* is negative and the minimum RPO is negative. If *bs* < *ad*, the cost of the lost data is decreasing faster than the cost of the system is increasing as RPO approaches zero. Therefore, the minimum cost point is at RPO = 0. If *bs* > *ad*, the cost of lost data is decreasing slower than the cost of the system is increasing as RP0 approaches zero. Therefore, there will be some RPO at which the total cost is minimum.



a=1.00, b=5.00, d=5.00, s=10.00, bs/ad=10

Minimum cost at RPO = 0.305

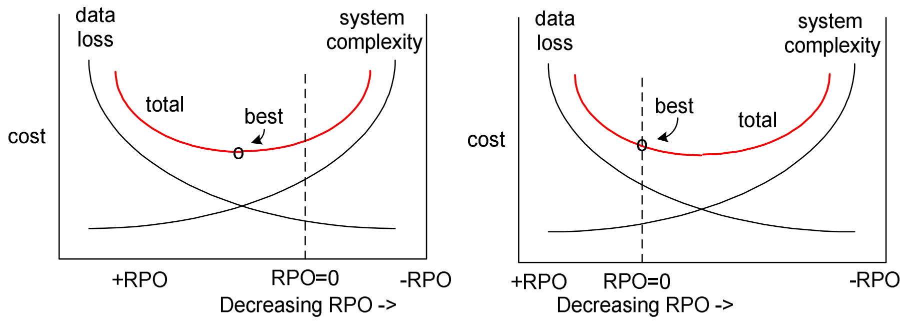
a=1.00, b=2.00, d=5.00, s=1.00, bs/ad=0.4

Minimum cost at RPO = 0

Exponential Cost Functions with Cost of Data Loss Increasing More Rapidly than RPO

Rule 2: *The cost of lost data must decrease slower than the increase in system cost as RPO approaches 0. Otherwise, the minimum cost point is at RPO=0.*

This rule is illustrated in the following graphic:

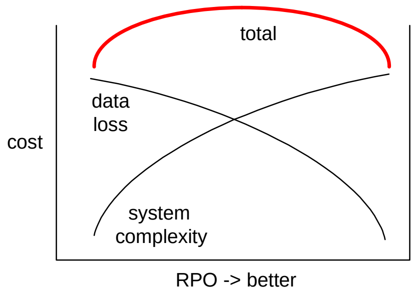


Data-loss Cost Decreases More Slowly than System Cost Increases

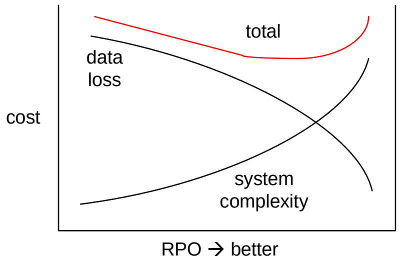
Data-loss Cost Decreases More Rapidly than System Cost Increases

Rule 2

Exponential functions with cost of data loss increasing less rapidly than RPO increases



In the previous case, both the lost-data cost and system cost curves were concave. If they had been convex, then the point of zero slope would be a total cost maximum rather than a minimum. If one curve is convex and one is concave, the point of zero slope could be either a maximum or a minimum.



We illustrate this with the case of exponentially changing data-loss cost and system cost in which the cost of lost data increases more slowly than RPO increases. That is, as RPO gets larger, the cost of lost data flattens out and increases ever more slowly. The cost of lost data and the system cost for this case can be represented by the relations:

$$\text{Cost of lost data} = d(1 - e^{-aRPO})$$
$$\text{Cost of system} = s(e^{-bRPO})$$

The total cost is then

$$\text{Total cost} = C = d(1 – e^{-aRPO}) + s(e^{-bRPO})$$

Setting the derivative of total cost to zero to find the minimum cost (or the maximum cost), we have

$$dC/dRPO = ad(e^{-aRPO}) –bs(e^{-bRPO}) = 0$$

$${e-}^{aRPO}/ e^{-bRPO} = bs/ad$$

$$e^{(b-a)RPO} = bs/ad$$

Solving for RPO:

$$(b-a)RPO – \ln(bs/ad)$$

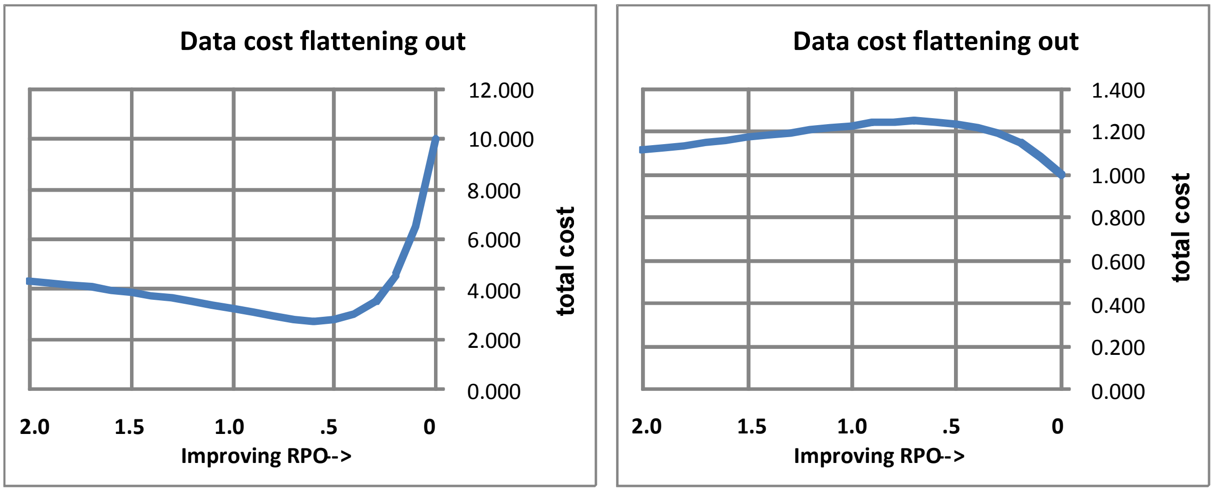
$$\text{minimum RPO} = \ln(bs/ad)/(b-a) \text{ for min or max}$$

This result also has an interesting interpretation. If *bs* < *ad* *or b* < *a*, but not both, the minimum RPO is negative. In this case, the best solution will be the system configuration for RPO = 0.

If *bs* > *ad* and *b* > a, there will be an RPO corresponding to a minimum total cost. However, if *bs* < *ad* and *b* < a (a negative divided by a negative), the point of zero slope of the total cost will represent a maximum cost, not a negative cost. The minimum cost will either be for RPO = 0 or for the maximum RPO.

bs > ad

bs < ad, b < a



a=1.00, b=5.00, d=5.00, s=10.00, bs/ad=10

a=2.00, b=1.00, d=1.00, s=1.00, bs/ad=0.5

Exponential Cost Functions with Cost of Data Loss Increasing Less Rapidly than RPO

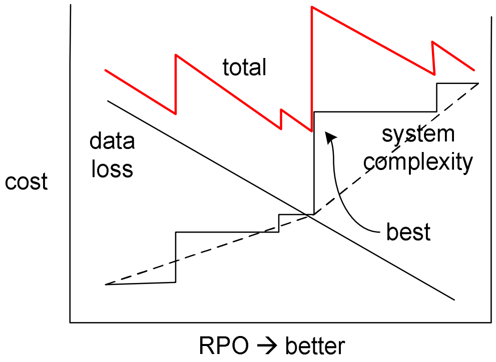
Example 3: Real-World Costs

A more accurate representation of the real world is:

- The cost of lost data decreases linearly as RPO approaches 0.
- The cost of the system increases in steps as one moves from one architecture to another more complex architecture to reduce RPO.

Therefore, the total cost is stepped as a move is made from one architecture to another.

The total cost can be determined graphically fairly simply:



- There will be many minimum total cost points
- Each minimum total cost will occur just before a system cost increase.
- Choose the absolute minimum.
- There may be more than one absolute minimum. In this case, choose the minimum with the smallest RPO.

Note that in the above graphical example, system cost is increasing slower that data-loss cost is decreasing up to the "best" RPO. Thereafter, system cost increases more rapidly than data loss is decreasing. See Rule 1.

Meeting an RTO

The cost analysis of meeting an RTO is similar to the analysis presented above for RPO. It is only necessary to replace "RPO" with "RTO" and the data-loss curves with downtime curves showing the increasing cost of downtime.

The system that minimizes RPO may not meet the required RTO. For instance, consider a system that must be "always up," but the data is not terribly important (such as a GPS application – if the system forgets where you are because it lost some data, it can quickly reconstruct where you are via another GPS reading).

Assume that the objectives are an RPO of four hours and an RTO of ten seconds. A virtual disk backup to a cold standby system might meet the RPO specification, but it would not meet the RTO specification. The RTO requirement might require asynchronous replication to a hot standby system.

Therefore, one must chose the least expensive system that meets both the RPO and the RTO objectives.

Rule 3: *The chosen system architecture must meet both the RPO and the RTO specifications.*

Summary

The proper system configuration to meet an RPO specification can be determined by plotting total cost as a function of RPO:

- As RPO is improved, the cost of data loss decreases and the cost of the system increases.
- So long as the cost of data loss decreases slower than the cost of the system increases as RPO approaches 0, there will be an optimal RPO to minimize cost.
- If the cost of data loss decreases faster than the cost of the system increases with improved RPO, the system providing a zero RPO will be the minimum cost system.

The optimum system to meet an RTO specification can be determined in the same way that RPO cost is analyzed. However, just because a system meets the RPO specification does not mean that it will meet the RTO specification. The system must be configured to meet both.

Acknowledgement

We would like to thank our subscriber, Dr. Bruce Holenstein, for suggesting this topic.