

NotesOn: Risk Management – Cumulative Recovery Time Objective

Introduction (v1.2):

A good friend and I were chatting the other day and one of those “everybody knows about” subjects came up, we were discussing Recovery Time Objective (RTO). During the conversation an heretofore “too obvious to mention” concept related to RTO crystallized into absolute clarity for me. You see, there is a flaw in the basic definition of RTO, an assumption that is built into it that can have serious negative effects on a business during and post any disaster event which requires recovery of IT systems.

Introduction (V1.2):	1
Background:	1
The Truth About RTO:	1
Cumulative RTO Definition:	3
Cumulative RTO Diagram:	3
Cumulative RTO (CRTO) Types:	4
CRTO Type 1:	4
CRTO Type 2:	4
CRTO Type 3:	5
CRTO Type 4:	5
Summary:	5

Background:

As you may recall from studying my post [Disaster Recovery & Business Continuity Definitions](#), RTO, or Recovery Time Objective, is an effort to quantify one of the recovery requirements of a business process and the IT system or systems behind it. From the above post: “[DRP] ... the desired chronological time from the beginning of a declared disaster event to when the application is ‘Ready for Use’ by the users.” In other words, if the system goes down, how soon does it have to be back up.

That seems simple enough, except there is a trap, a camouflaged hole, built into the definition if taken literally, or ... if presented to the business users without providing a full understanding of what RTO implies.

The Truth About RTO:

The truth about RTO is that there is more to it than what appears on the surface of the core definition.

As a rule, when most IT-to-business representatives (i.e. embedded business analysts, IT system stewards, etc.) are asked for the RTO value for System A, when most if not all business users are asked for their RTO requirements for System A, they respond in terms of “minimum discomfort level”, i.e. after such and such a



time without System A (1 hour, 24 hours, 3 days, and so on) the business process(es) associated with System A becomes “a pain”.

When describing their RTO requirements these folks think in terms of the recovery of only *that* system. Unless they have lived through a “full blown” IT disaster recovery cycle, a component failure of that system, which briefly disrupts normal routines, is all they have experienced.

However.

When discussing RTO we are not addressing normal system maintenance and support issues but, rather, disasters. There is a world of difference between a singular event taking down one system and a disaster event taking out, in one way or another, multiple systems, as in many to all of a Business’s systems.

Where that difference is not recognized, if the business users are not made aware of the true range of potential recovery times, if they have not been allowed to think outside the “normal occurrences” box, if they have not been reasonably educated on *all* of the actions and activities and time elements inherent to IT Disaster Recovery Plans ... when the time comes they will be sorely disappointed in IT’s response to their specified RTO (and, of course, the delay in the re-automation of their business processes).

Why? Because without a full understanding of RTO, the Business Unit’s Business Continuity Plans have likely been tainted. In all probability the Business Continuity team, in accepting the “ideal” RTO value, will not have mapped all of the manual processes and procedures and resources required to cover the RTO gap.

For example, if the desired RTO was stated as 24 hours and “everyone” planned on 24 hours but, post a full on Katrina Hurricane type disaster it factually takes two months (or longer) to bring their system(s) back up ...

As I’ve mentioned elsewhere, a close knit partnership must exist between IT and its Business. This is never more true than when discussing RTO. A balance must be struck between what IT can do, and afford to do, and what the Business Unit (BU) desires to have and what it absolutely needs to have to survive. Understanding this delta, this difference lays the groundwork for addressing:

1. The necessary, workable, DR architecture of the target system, *and* its supporting systems;
2. The IT budget to pay for appropriate failover of the target system, *and* its supporting systems;
3. The realistic Disaster Recovery Plan for the target system, *and* for its supporting systems;
4. The contingency (business continuity) plan for the Business Unit that spans a potential RTO gap that ranges from “simple” outages to complex systems failures to outright datacenter losses;
5. The Emergency Operations Center’s initial strategic and tactical plans so they are not overly optimistic and thus of little value in a true disaster situation.

Cumulative RTO Definition:

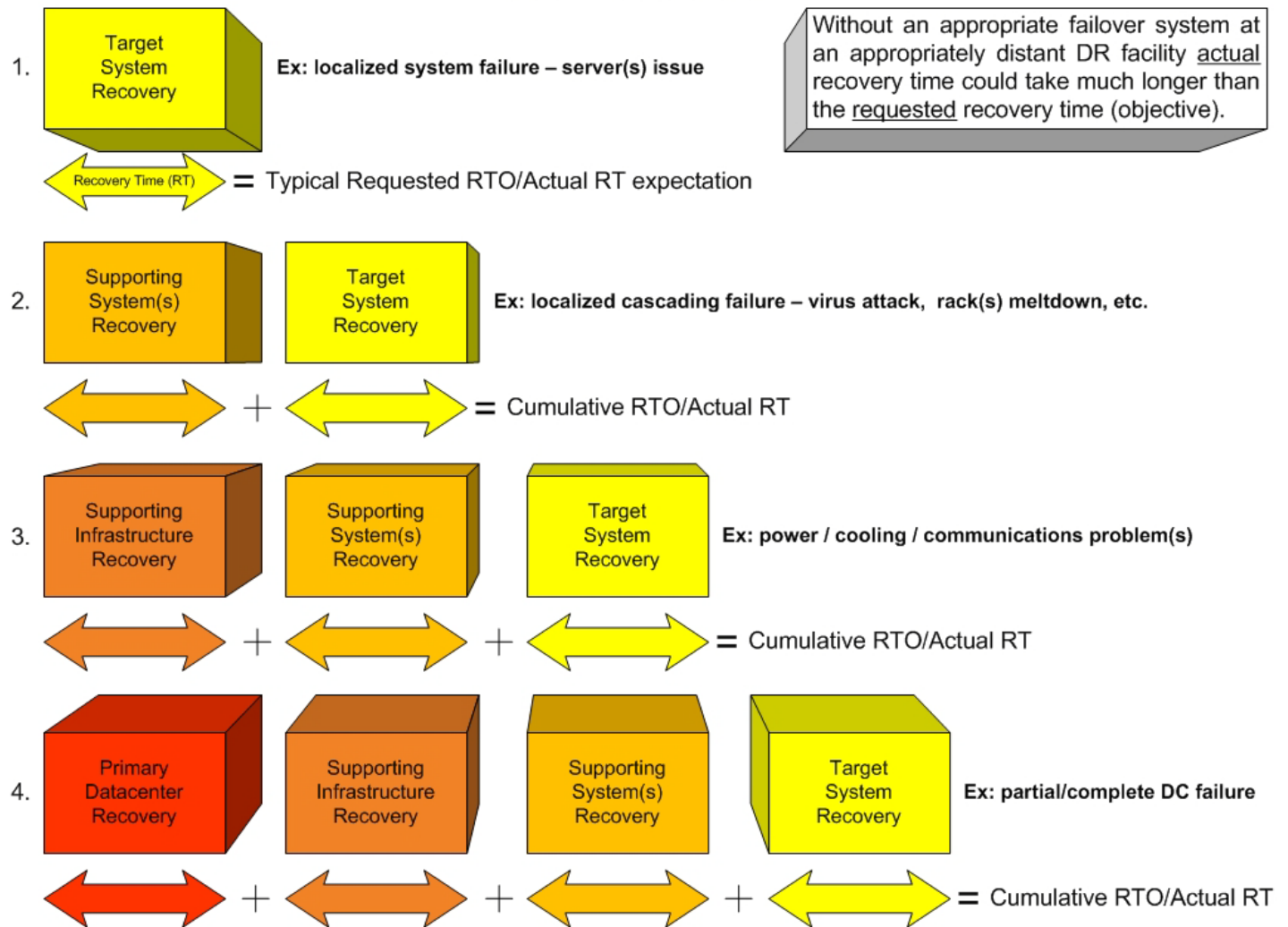
Cumulative RTO: the accumulating effect on a target system’s Recovery Time Objective (and actual Recovery Time) if one or more layers below the target system also goes down. The layers being considered are: (1) the target system (2) one or more supporting systems -- systems upon which the target system depends (3) the supporting infrastructure – networking, communications, ETL utilities, etc. and (4) the primary datacenter in which the target system resides. Failure of more than the target system will result in an extended outage that will impact the typically expressed RTO value. Example: the requested RTO for System A is 24 hours but, in the absence of any failover systems and environments, if Layers 2, 3 and/or 4 are also down the requested RTO will not be met, unless it knowingly took into account the cumulative recovery time of all layers.

Cumulative RTO Diagram:

Please refer to this diagram as we continue our discussion on Cumulative Recovery Time Objectives.

Business Continuity/Disaster Recovery: Cumulative RTO Diagram – v1.1

© DP Harshman - www.fromtheranks.com



Cumulative RTO (CRTO) Types:

Take a minute or two to review the above diagram. If you haven't already, you may very well have one of those "Ahhh-hahhh!" moments as the clarity of the "Cumulative RTO" concept emerges. It truly is this simple. But. Again. As with other DR and BC subjects, its simplicity has kept CRTO as a fully formed concept off the radar and thus never called out in a manner easily passed on to others, especially non-IT business users.

Though it *is* straightforward let's review each of the, for lack of a better term, "Cumulative RTO (CRTO) Types".

CRTO Type 1:

This is the type that most everyone thinks of when discussing and setting RTO values. "System down, call IT, system up, thank you, now let's all get back to work and catch up as quickly as we can". Type 1 is totally system-centric. An element of that system, often hardware (as in a database or file storage or application server) but sometimes software (the website itself, the database, etc.) goes down. If the unintentional outage runs overly long then a system failover occurs, assuming there is a failover system. If there is no failover plan then the system comes up when it comes up but usually IT is pretty good about these recoveries. Even in worst cases where a new server has to be bought and brought in, IT can usually have it up and running within a day or two. Unfortunately Type 1 does not describe the typical disaster event scenario.

CRTO Type 2:

With a Type 2 the situation escalates, rapidly. Events that describe this scenario include virus attacks that "wipe out" the entire target system *and* one or more of its supporting systems – by supporting system I am referring to systems that feed the target system with critical information and/or services without which the target system cannot properly function.

Or, it could be that an electrical fire takes out an entire "rack" of servers, frying the hardware – unlike a typical desktop configuration, in a datacenter the hardware is stacked in six to eight foot tall racks that allow easy connection to power and network cables as well as air conditioning.

This is not the type of event where the National Guard is called out, or the Emergency Operations Center (EOC) is typically activated. It is, however, the type of event that raises the frustration level of the affected business unit(s) because, unless there are true failover systems somewhere else, and unless the BU's truly understood RTO and Cumulative RTO during their interviews, the RTO they envisioned (CRTO Type1) isn't being met and the budget spend they invested in IT isn't buying them what they expected.

A CRTO Type 2 incident may require affected BU's to activate their Business Continuity Plan but only so far as gearing up the manual procedures that fill the gap created by the system going down hard. In the meantime of doing that they are, of course, on the phone to IT and IT Management, regularly, demanding to know when their critical systems will be back up. They will be upset because they can't get their work done and, since they were depending on CRTO Type 1 recovery times, because they haven't made adequate provisions for manual work-a-rounds.

CRTO Type 3:

Type 3 is the next worst type of incident for IT and for the Business. The “bricks and mortar” of the primary datacenter are intact but, to reach Type 3 status, a major “problem” has occurred in the datacenter’s environment. An example might be a heat wave that overpowers the electrical grid supporting the datacenter forcing the DC onto its backup generators, which in turn “cook” due to their inability to “throw off” the excess heat. Another scenario might be an ice storm taking out power lines and closing roads and highways. The backup generators keep the systems running for a while but with the fuel supplier unable to get out to the datacenter the diesel tanks will eventually run dry, taking everything in the datacenter down.

Once the “incident” is over the datacenter will gradually bring everything back on line, starting with their supporting infrastructure – external and internal network connections, routers and switches, etc. – and gradually working up to the point where IT (now that it can connect to the datacenter) begins spinning up supporting systems that feed each Tier 1 target system, eventually getting to the Tier 1 systems themselves.

With Type 3’s the National Guard may very well have been called out in aid of “life support” activities and the EOC may very well have been activated. The EOC in turn initiates each BU’s Business Continuity Plan (BCP) which, hopefully, has manual procedures to compensate for the displaced automated systems. Unless ... there is a DR facility in a location outside of the DR event zone to which IT has already failed over at least the critical systems ... and those upon which they depend.

CRTO Type 4:

CRTO Type 4’s are the worst possible nightmare come to life for a Business. Everything from the brick and mortar of the datacenter to all of the systems within it are not only down but ... gone. Power and phone grids around it may be gone as well. To be intentionally disturbing, without a DR facility having been set up outside of the primary datacenter’s event zone, the Business had better have business continuity plans that address *long* term recovery times, including the manual procedures to back up the orphaned business functions. The cumulative RTO for this type of event is: primary datacenter recovery (rebuild) plus supporting infrastructure recovery (rebuild) plus supporting systems recovery (rebuild) plus Tier 1 systems recovery (rebuild).

Summary:

Without a true and complete understanding of all of the aspects of RTO, both IT and the Business will be very dissatisfied with IT’s disaster recovery efforts. If the Business *truly* understands that the loss of a datacenter, or supporting infrastructure, or supporting systems will have a direct impact on when their critical systems come back up they will be far more supportive of IT’s efforts and most likely more supportive of a Disaster Recovery facility, as well as investing in both IT’s Disaster Recovery Plans and their Business Continuity Plans.

Hope this helps,

DP Harshman

PDF Link