Incident Management Procedures

Version 1.2 04/08/2011

Introduction

This document outlines in broad terms the procedures to be followed in restoring a service which has become unavailable for any reason. This document does not in general differentiate between losses of service due to accident or hardware failure and losses of service due to malicious action. If the latter is suspected, then additional actions as laid down in the compromised machine policy may need to be taken. The procedures detailed here for restoring corrupt data can be applied regardless of whether the corruption is deliberate or not with the proviso that there may be a need to preserve corrupted data as evidence if a compromise is suspected.

A Brief Plan of Action

The following basic steps should be followed in the management of any major incident:

1. Inform users
2. restore hardware
3. restore data
4. report including any lessons learned

Not all of these steps will necessarily apply for every incident.

Communication With Users

The most important aspect of restoration of a service is to make sure that users affected by the service disruption are kept informed at all stages of the process. The communication method will depend on the number of users involved and on the nature of the incident. Email will normally be the most efficient method of communication, either in the shape of emails to individuals if the number of users affected is small, or via more general broadcast methods such as the sys-announce mailing list, but there may well be occasions where email is not available and in this case alternative channels of communication must be found. Some alternative methods of communication are the status message on the computing support front page, a flip chart located at the main entrance of the School's buildings displaying an up-to-date summary of progress and even members of the support team going from office to office updating users verbally.

From the above, it will be seen that it is absolutely essential for the members of the User Support Unit to be kept fully up to date with progress so that they can respond quickly and accurately to user's queries. For significant incidents (those likely to last more than two hours or so), both the User Support Unit and the Unit primarily dealing with the incident should designate people to act as incident coordinators. These coordinators should communicate with each other at least hourly even if the only information exchanged is that no progress has bee made. The US coordinators should then disseminate this information to the US unit and beyond.

When communicating with users, care should be taken to not overload them with information. If desired, additional details can be provided via alternative channels of communication such as the system blog. Users should be informed about the issue as soon as possible and should be provided with an estimate of the length of time needed to restore the service as soon as this can be reasonably made. Should the incident last for more that 24 hours, affected users should be informed at least daily of progress until the incident is resolved.

Prerequisites for service recovery

To facilitate the prompt restoration of services, the School should always have to hand spare hardware which can be used to replace failed hardware which cannot be restored to operation within 4 hours or so. This should include reasonable substitutes for every model of server hardware used (in most cases it will probably be sufficient to use a similar specification of machine rather than an identical model) and replacements for individual hardware items such as Fibre HBAs, network cards and SCSI cards. These items should be in a well publicised location accessible by all members of the computing staff. In addition, the Services Unit should ensure that there is always a significant amount of disk space (of the order of 10TB to cater for the loss of an entire array) online and available for data restoration. For services which are site dependant (for example services which make use of data on a site's SAN), suitable replacement hardware should be available at the same site.

Consideration should also be given to replacing failed servers with virtual machines where this is not ruled out by physical restraints such as a requirement for attachment to one of the School's SANs.

Initial steps

When restoring a service, it is important to strike a balance between the understandable desire to restore normal service as quickly as possible and the vital need to ensure that the integrity of any data associated with the service is still intact. Since each incident will be different, the final decision on the amount of time to spend on investigation will come down to the experience and knowledge of the person managing the incident. As a very general rule however, should there be any suggestion that data corruption has occurred, this possibility should be investigated exhaustively before restoration of the service begins.

It may be useful to stage the recover of the service since this will allow investigators to monitor the service for signs of repeated failure or corruption without allowing the general user base access to the service. This might be achieved by using service specific configurations to restrict access to certain users or IP addresses, more general tools such as tcp wrappers or, in the case of an externally accessible service, by not opening the firewall hole to the outside world until the integrity of the service has been satisfactorily established.

Restoration of hardware

In many cases, a simple reboot will be sufficient to restore the service. In that case, the investigation can begin immediately. Should any potentially repeatable issues be found, it is vital to inform the relevant Unit so that action to prevent a re-occurrence can be taken.

If a reboot does not fix the problem, attempt to ascertain whether the issue lies with software or hardware. If the problem is clearly caused by hardware failure the issue may possibly be resolved by swapping components from the backup hardware, for example PSUs, HBAs, Network cards and RAM. Failing that, the backup hardware should be brought into service as quickly as possible. To this end it may be worth the extra energy cost to maintain the backup servers as hot spares allowing them to pick up OS upgrades etc. The sleep component could be used on these servers to minimise energy costs. Care should be taken however to ensure that these machine are not merely treated as another resource to be used at will compromising their ability to substitute for failed hardware at short notice.

If the failure can be clearly established to be software related, members of the appropriate Units should be called upon to give the benefits of their expertise. Efforts should concentrate on determining what has changed to cause the problem with full consideration being given to reverting to earlier versions of the software if the issue cannot be resolved quickly. For this reason, it is always desirable that software upgrades should have a clearly defined back-out path.

If it is not clears whether the problem is software or hardware related, the most effective path may well be to divide up effort with one group concentrating on the software aspect and another bringing the backup hardware online. Seeing whether the problem appears on the new hardware or not may give a clear indication of where the problem lies.

It's good practice when setting up a service which is associated with an IP address to make that IP address different to the main address of the server. This will considerably ease the task of moving the service to new hardware.

If a service is restored by relocating it to different hardware, it is important to ensure that the original hardware does not come back at a later date and start offering the same service, particularly on the same IP address as the current version of the service is using.

Restoration of Data

Once the service's hardware is available again, it must be decided whether it is necessary to restore some or all of the data associated with the service. Sometimes this decision will be straightforward (for example where the underlying storage has failed with no hope of recovery) but at other times the issue will be less clear cut. A prime example of this is the situation where the service data is still available but there is the possibility that the data has been corrupted or compromised. It is therefore vital to establish whether data

corruption or compromise has taken place as this will dictate both the need to restore data and the procedure to follow for the restoration.

If there is no likelihood of data corruption, then the restoration procedure is straightforward, copy the data from the mirrors (most up to date), most recent backups or golden copy (for data which is not backed up) to the replacement or repaired storage hardware. The data will probably be split into chunks which can be brought online individually saving time for some users at least. It may even be possible to make the mirror data directly accessible saving time in restoring the service but complicating the clean up procedure afterwards.

Possibly corrupted or compromised data will need a more painstaking approach. The integrity of the data may be suspect if users have been reporting issues with data, service specific or more generalised tools (such as fsck in the case of file systems) report inconsistencies in the data or unexpected executable and particularly setuid files appear in the data (the latter strongly pointing to the possibility that the data has been compromised by malicious action). The recommended course of action to take in the case of suspected corrupted data is as follows:

1. Do NOT make the suspect data available to the users. The users should however be informed of the suspicion that corruption has occurred.
2. Determine the last date at which the data was known to be good. This may be done by looking back through user reports of problems, system logs etc. Restore the backup of the data from this point or if backups do not exist, reinstall the data from the golden copy. Do NOT make this data available to the users either.
3. Compare the latest and last known good data using appropriate tools. If the two sets of data are identical, make the data available to the user base. If there are differences and these changes are not obviously the result of corruption/compromise, encapsulate these somehow in a user understandable format (the details of how this is done will obviously depend on the type of data) and ask the owner of the data to confirm that these changes are expected. If they can do so, make the latest data available to the user base. If the user cannot confirm that these are expected changes or the data is obviously corrupted/compromised, make the last known good data available and determine as far as possible what changes were made to the data since the last known good backup.

Lessons Learned

Once the incident has been successfully managed, time should be taken to write a short report covering the causes of the incident, the procedure followed to restore the service, whether with the benefit of hindsight things could have been managed better and steps that should be taken to prevent the incident re-occurring. Any other lessons learned should also be included in the report. The system blog may be a suitable medium for this report and it may also be worth raising it as a topic for discussion at an Operational meeting.

Last Words

Above all, remember to keep the users informed at the beginning, during and at the end of

the incident.