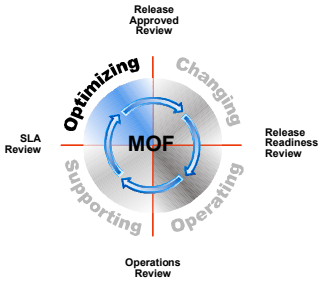
Microsoft®

MOF Service Management Function

Availability Management

patterns & practices



Microsoft®

Solutions for Management

Contents

Document Purpose

This guide provides detailed information about the availability management service management function (SMF) for organizations that have deployed, or are considering deploying, Microsoft technologies in a data center or other type of enterprise computing environment. This is one of the more than 20 SMFs defined and described in Microsoft® Operations Framework (MOF). The guide assumes that the reader is familiar with the intent, background, and fundamental concepts of MOF as well as the Microsoft technologies discussed.

An overview of MOF and its companion, Microsoft Solutions Framework (MSF), is available in the *Introduction to Service Management Functions* guide. This overview guide also provides abstracts of each of the service management functions defined within MOF. Detailed information about the concepts and principles of each of the frameworks is also available in technical papers available at www.microsoft.com/solutions/msm.

Executive Summary

Availability has become one of the most important aspects of service delivery in the highly visible e-business global economy. Consequently, the demand for 24-hours-a-day, 7-days-a-week operation is greater than ever. Availability, or the lack of it, has a dramatic influence on customer satisfaction and can very quickly impact the overall reputation and success of the enterprise. Availability management is responsible for ensuring that service-affecting incidents do not occur, or that timely and effective action is taken when they do.

Risks to availability may be caused by technology, processes and procedures, and human error. Countermeasures, such as carefully designed testing and release procedures and appropriate staff training plans, can be employed to help mitigate these risks. Risks to availability exist throughout the whole IT infrastructure and within every management process. Although not directly responsible for each of these processes, availability management is responsible for making sure that all areas of risk to availability are taken into account and that the overall IT infrastructure and the maturity of management processes supporting a given IT service are sufficient.

Availability management and service continuity management are closely related in this respect as both processes strive to eliminate risks to the availability of IT services. The prime focus of availability management is handling the routine risks to availability that can be reasonably expected to occur on a day-to-day basis. Rare, expensive, or unanticipated risks are handled by service continuity management.

Process and Activities

Overview

The availability management process focuses on two distinct areas, although many of the overall tasks are the same. The first is a new IT service that clearly involves the fullest requirements definitions and design phases. The second area is existing IT services that may require significant short-term effort to improve levels of availability.

New IT Services

New IT services provide the best opportunity for achieving availability targets in a cost-effective manner because availability considerations can be built in from the earliest stages. This allows the most appropriate technologies to be selected and an IT support infrastructure to be built that provides the required level of operational maturity.

The customer and IT organization have the best opportunity in this scenario to work closely together on the definition and level of availability to be provided by the IT service and to agree upon the level of investment required. This avoids inappropriate expectations from emerging in the future and allows any mismatches between the levels of availability and the investment required to be resolved early.

The aim of new IT services is to achieve the desired availability targets from day one and to successfully manage the levels of availability throughout the life cycle of the solution. It is particularly important to manage levels of availability during the introduction of the functional and technological changes demanded by today's fast-moving business environments.

Existing IT Services

Existing IT services can have their availability levels significantly improved or stabilized through the adoption of a formal availability management process. They can then benefit from an ongoing continuous improvement process and careful management of future changes.

The challenge of improving availability levels in existing IT services is they often come with a legacy of design constraints and technology challenges that may not be cost-effective to overcome. That is why building availability in from the very beginning is so important.

The life cycle approach is very similar to new IT services and begins with a definition of availability with the customer and determination of an appropriate budget for improvements and ongoing maintenance that can be justified by the cost of downtime.

In some respects, existing IT services have an advantage over new IT services in that they have a track record of service delivery that can be examined in detail and any shortcomings and areas of exposure addressed. The availability design process includes an investigation of the history of service outage experienced by the customer, as well as root cause analysis as appropriate.

At any one time, the role of the availability manager invariably includes both the improvement of existing IT services as well as the introduction of new IT services. In addition, the introduction of major change to an existing IT service, such as upgrading from one technology platform to another, also involves mixing these two scenarios. The basic process to be followed is the same in either case.

Goals and Objectives

The objective of the availability management function is to ensure that any given IT service consistently and cost-effectively delivers the level of availability required by the customer.

Scope

Availability management is concerned with the design, implementation, measurement, and management of IT infrastructure availability to ensure that stated business requirements for availability are consistently met. In particular

- Availability management should be applied to all new IT services and for established services where service level requirements (SLRs) or service level agreements (SLAs) are established.
- Availability management can be applied to IT services that are defined as critical business functions, even when no SLA exists.
- Availability management can be applied to the suppliers (internal and external) that form the IT support organization as a precursor to the creation of a formal SLA.
- Availability management considers all aspects of the IT infrastructure and supporting organization that may impact availability, including training, skills, policy, process effectiveness, procedures, and tools.
- Availability management is not responsible for Business Continuity Management and the resumption of business processing after a major disaster. This is the responsibility of the service continuity management SMF. However, availability management is closely related and provides key inputs to service continuity management.

Key Definitions

The following are key definitions within the availability management processes:

Availability. Ability of a component or service to perform its required function at a stated instant or over a stated period of time.

Countermeasures. Actions taken to prevent or reduce the effect of an identified risk.

Critical business functions. The critical elements of the business process supported by an IT service.

Downtime. The unavailability of the IT Service during hours that the business deems the systems to be available—as advertised within SLAs.

End-to-end service. All components of the IT Infrastructure required for delivering an IT service.

High availability. Minimizing or masking component failures.

Incident life cycle. An availability technique which analyses the broken down stages of an incident to allow for timing and measurement of each stage.

Maintainability. The ability of an IT infrastructure component to be retained in, or restored to, an operational state.

Operating level agreement. An internal agreement covering the delivery of services that support the IT service provider in the delivery of services.

Risk Management. The identification, selection, and implementation of countermeasures to the identified risks to assets to reduce them to an acceptable level.

Reliability. The freedom from failure of services and components over a given period of time.

Serviceability. The contractual arrangements made with Third Party IT service providers to provided or maintain IT Services or components.

Service level agreement. Written agreement between a service provider and the customer(s) that documents agreed service levels for a service.

Service outages. See downtime.

Major Processes

Availability Management comprises of three main processes and a number of subprocesses as follows:

- Define service level requirements
- Define critical customer functions
- Define availability objectives
- Propose availability solution
- Identify major Information Technology service components
- Design for availability
- Availability risks and countermeasures
- Life cycle management needs
- Design for recovery
- Incident life cycle
- Designing for customer satisfaction during outages
- Management processes
- Formalize operating level agreements

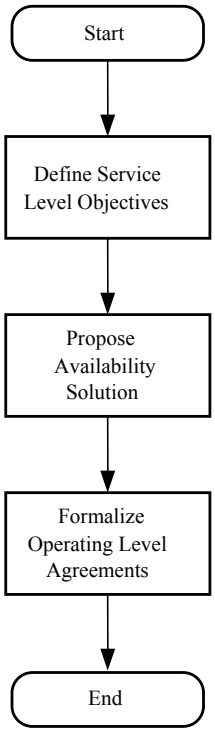


Figure 1
Availability process flow diagram

Define Service Level Requirements

The terms availability and "high availability" can have very different meanings depending on the context in which they are used and the audience involved. They can be used to describe a wide range of business goals and technical requirements, from relatively easily achieved hardware-only availability targets, to very demanding mission-critical targets applying to the availability of the IT service as a whole.

As a result, it is relatively easy for inappropriate expectations to be set concerning availability targets. It is also easy for the customer community to demand higher levels of availability than they are actually willing to pay for once the cost implications are understood.

Availability management needs to start with carefully agreeing to availability targets with the customer and determining the cost of downtime or unavailability of the IT service in question so that a realistic IT budget can be established.

This process involves an element of education and negotiation on both sides as the customer needs to understand how to define and articulate their availability requirements. The IT organization needs to understand the different customer functions that make up the overall IT service and which of these are the most critical.

The overall task of negotiating service level objects is defined in the service level management process. A brief overview of the process is presented here to provide a background. For more information, see the MOF service level management operations guide.

Define Critical Customer Functions

Any given IT service often contains multiple customer functions or transaction, and some of these will have varying availability requirements and impacts on the business if they fail. Each of the major business functions or transactions that make up the overall IT service needs to be identified and ranked in order of importance with the most critical elements being singled out for particular attention.

Interdependencies between services and any reliance on less-important services also need to be identified. For example, a new critical call-handling system may rely heavily on file and print services being provided from elsewhere within the IT infrastructure, and this may require the criticality of the file and print service to be upgraded accordingly.

Define Availability Requirements

The availability of an IT service is a complex issue that spans many disciplines. There are many different IT approaches that can be taken to deliver the required levels of availability, each with their own cost implications.

In complete contrast to this, however, availability requirements can often be expressed in relatively simplistic terms by the customer and without a full understanding of the implications. This can lead to misunderstandings between the customer and the IT organization, inappropriate levels of investment, and ultimately to customer dissatisfaction through inappropriate expectations being set.

One expressed requirement for 99.5 percent availability can be completely different to another requirement for 99.5 percent. One requirement may discuss the availability of the hardware platform alone, and the other the availability of the complete end-to-end service. Even the definition of complete end-to-end service availability can vary greatly, and it is important to understand exactly how any availability requirements are to be measured. For example, consider:

- If all hardware and software on the primary server is functioning correctly and user connections are ready to be accepted by the application, the solution might be considered 100 percent available.
- If there are 100 users but 25 percent are unable to connect due to a local network failure, is the solution still considered 100 percent available?
- If only one user out of the 100 is able to connect and process work, is it only 1 percent available?
- If all 100 users are able to connect but the service is degraded with only two out of three customer transactions being available, or there is very poor performance, how does this affect availability measurements?

The period over which availability is to be measured can also have significant impact on the definition of availability. A requirement for 99.9 percent availability over a one-year period allows 8.8 hours of downtime. A requirement for 99.9 percent availability over a rolling four-week window only allows 40 minutes downtime in each period.

The customer should define availability in his or her own terms. The availability manager should strive to educate and make sure that he or she fully understands the terminology and that the end result is realistic. The customer takes more ownership for the consequences when they have been allowed to determine the measurement criteria.

It is also necessary to identify and negotiate periods of downtime for planned maintenance activity, technology upgrades, and the introduction of new business function. The amount of planned downtime that can be tolerated by the customer again has a significant impact on the definition of availability requirements.

Propose Availability Solution

Now that the availability objectives for the critical business functions within an IT service have been identified and their relative importance and financial implications clearly understood, the task of designing an overall IT infrastructure that delivers the required levels of availability can begin.

This overall design process is not unlike traditional development methodologies. Availability management is part of the optimizing phase of MOF but often overlaps with the duties performed in the envisioning phase of MSF. Availability management is responsible for setting appropriate, availability-related requirements and standards so that a formal project, managed through MSF, can ensure that IT services are built as required.

The availability design process often involves a number of iterations between the preceding requirements definition phase and the analysis and design phase. The availability goals and budgetary constraints defined during the requirements phase may prove to be unachievable and need to be passed back for re-negotiation with the customer.

The first step in the availability design process is to identify the major IT technology components, infrastructure, people, and processes that underpin the complete end-to-end delivery of service for each of the critical business functions and transactions to be protected.

The complete life cycle of each of these components can then be considered in detail with a view to designing a highly available IT infrastructure and support environment. There are then two main design processes to be undertaken: designing for availability and designing for recovery.

Designing for availability can be thought of as a pro-active task where the focus is on keeping the IT service running and preventing service outages from occurring or reducing the impact of failures on the service being provided.

Designing for recovery is predominantly a re-active task that aims to diagnose and recover service as quickly as possible if it actually fails or becomes degraded in some manner.

Identify Major Information Technology Service Components

During the first step in the availability design process, it is useful to break down the end-to-end provision of any service into manageable pieces and to examine each of these pieces in turn. MOF breaks down the overall infrastructure that makes up and supports an IT service into the following IT domains:

- Service
- Application
- Middleware
- Operating system
- Hardware
- Network
- Facilities
- Egress

Design for Availability

The complete life cycle of each IT component identified above can now be considered in detail with a view to maximizing the availability that is delivered by it.

The appropriateness, reliability, maintainability, and serviceability of each IT component can now be examined in detail and considered from two major perspectives:

- Availability risks and countermeasures.
- Life cycle management needs.

Availability Risks and Countermeasures

All risks to the availability of each IT component need to be considered and appropriate countermeasures designed to mitigate them.

The nature of the availability risks faced by an IT component varies according to the MOF IT domain the component resides in.

Examples of availability risks by IT domain are:

- Application, middleware, and operating system domains:
- Single point of failure.
- Incorrect configuration option.
- Design flaw.
- Poor development methodology.
- Coding error.
- Hardware and network domains:
- Single point of failure.
- Out of date firmware.
- Poor documentation.
- Vendor support quality.
- Lack of anti-static precautions.
- Lack of spares.
- Poorly labeled cabling.
- Facilities domain:
- Insufficient air-conditioning capacity.
- Power outages.
- Power surges and spikes.
- Fire and flood.
- Physical security.
- Egress domain:
- Single power feed from utility.
- Single communications feed from Telco.
- Personnel:
- Poor quality procedures.
- Lack of discipline.
- Lack of skills.

Availability management and service continuity management are closely related. Both processes strive to eliminate risks to the availability of IT services and employ the use of countermeasures to achieve this. The prime focus of availability management is in handling the routine risks to availability that can be reasonably expected to occur on a day-to-day basis. Service continuity management caters to more extreme and relatively rare availability risks, such as fire and flood, and also acts as a catchall for any unanticipated availability risks.

Service level management impacts both service continuity management and availability management. Service level management takes primary responsibility for interfacing with customers and determining which IT services are most crucial to the survival of the company, and which alternate means of conducting business are employed if they fail for a prolonged period.

Availability management draws on this prioritization work and takes it a stage further by identifying the key IT infrastructure components that support these critical services and determining whether they contain any single points of failure or other risks to availability that can be cost-effectively addressed through the use of appropriate countermeasures.

Where no straight-forward countermeasures are available or where the countermeasure is prohibitively expensive or beyond the scope of a single IT service to justify in its own right, then these availability risks are passed to service continuity management to handle.

Within each IT domain, there are specific risks to availability that are considered too unlikely to justify the cost of mitigating them, or there are risks that were not anticipated. For example, few data centers anticipate a meteor shower on the building. Of those that do, few spend the money to install anti-meteor shielding. In these situations, service continuity management outlines what must be done to restore service. There need not be a separate service continuity plan for each risk. One plan can cover the risks of flood, fire, meteors, terrorist attack, and any other eventuality that might disable a complete data center. There always needs to be a service continuity plan in place, even where there is also an availability plan in place to handle more routine issues.

As with an earlier example, an individual power supply within a server can be expected to fail at some stage. A very effective and inexpensive countermeasure that can be employed by availability management is the adoption of server technology incorporating hot-plug redundant power supplies. This technology allows a second power supply to seamlessly take over from the failing unit and for the failed unit to be replaced on-line without any interruption to the IT service. Service continuity management needs to plan for situations where both power supplies fail at the same time, or where the second fails while the first is being repaired. This is a much more unlikely scenario, but must be planned for.

The following figure summarizes the relationship between availability management and service continuity management with regards to the identification and mitigation of availability risks.

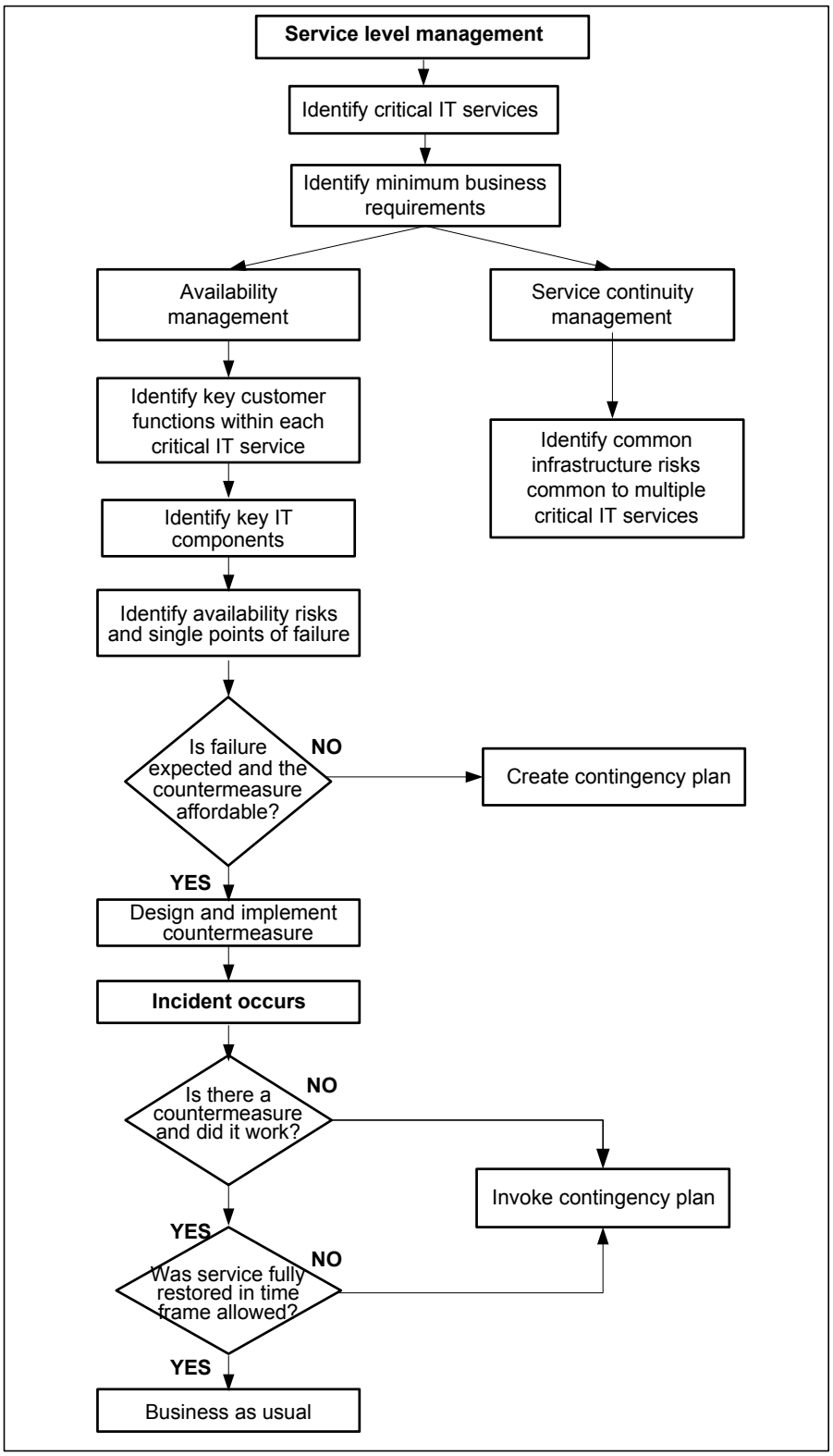


Figure 2

Relationship between availability management and service continuity management

When an availability risk is identified and confirmed as falling within the scope of availability management, the next step is to identify an appropriate countermeasure that can be deployed to minimize the exposure to the IT service.

It is important to ensure that any countermeasures employed are affordable and can be cost-justified in relation to the cost of downtime that has already been agreed with the customer. Availability management strives to provide an optimum level of availability within the cost constraints imposed by the customer.

A countermeasure may be only partially effective by design. For example, a stand-by system may have only half the processing capacity of the primary system. Whether this is acceptable depends on the definition of availability agreed upon with the customer.

Where a particular risk cannot be addressed at an appropriate cost, then one of the following must occur:

- The availability goals need to be re-negotiated with the customer.
- A strategy of rapid-recovery needs to be adopted (with its associated impact on availability levels).
- The issue must be passed over to service continuity management and any resulting outage handled as an exception/disaster.

Although service continuity management provides contingency plans to handle any disaster, these may involve a prolonged period of downtime before the IT service is fully restored. This factor needs to be taken into consideration during re-negotiation with the customer.

Such iterative cycles of customer need, risk identification, and design implication may need to be followed several times during the design and implementation of a highly available IT service.

Life Cycle Management Needs

The full life cycle management requirements for maximum availability of each IT component must be determined. Then, the appropriate roles and responsibilities, tools, processes, and procedures can be designed into the supporting IT infrastructure along with the necessary staffing levels and skill requirements to carry them out.

The aim is to identify any operational task that can be undertaken to maximize the availability of the IT component, such as fast startup and shutdown times, and attending to housekeeping needs in a timely fashion. Opportunities for performing periodic health-checks and other pro-active monitoring must be explored and the need for custom instrumentation or specific tooling considered.

Such management needs for each IT component includes efficient handling of:

- Startup and shutdown, including subsystem dependencies.
- Monitoring.
- End-to-end health-checks.
- Housekeeping/administration.
- Password maintenance.
- Consumable replacement methodology.
- Backup and restore methodologies.
- Emergency patch methodology.
- Upgrade and change methodology.
- Opportunities to exploit on-line backup and on-line configuration features.
- Fail-over and fail-back requirements.
- Skills required to operate, monitor, and diagnose.
- Event generation and handling requirements.
- Configuration documentation required.
- Up-to-date vendor documentation.

To maximize the availability of the IT service, management processes and an IT infrastructure that support the operational needs of underlying IT components must be designed. Such a design encompasses the roles and responsibilities of the various groups involved, the tools and techniques they employ, and the specific tasks both real-time and off-line required to keep the IT component operational. The design must also be able to deal with both routine expected problems and failure scenarios. Any requirements for invasive maintenance needs to be scheduled in accordance with the agreed service windows and included within internal and external contracts and SLAs.

Design for Recovery

No matter how well designed and managed, problems with the delivery of an IT service can still occur. The second major design consideration for high availability is a reactive one charged with recovering service as quickly and efficiently as possible. The incident being dealt with may be the result of an unexpected event or even the failure of a countermeasure to protect the service. Rapid recovery may also be the appropriate design choice for a particular availability risk if an effective countermeasure proves to be too expensive for the customer to justify.

Efficient incident detection and recovery mechanisms are also required during less extreme situations, as even minor problems need to be handled appropriately to prevent knock-on errors from escalating further down the chain. In the case of dual redundant power supplies, any failed primary component clearly needs to be identified and replaced before the secondary unit also fails and results in total loss of service.

As illustrated in the following figure, every incident moves through these life cycle stages:

- Incident start
- Incident detection
- Incident diagnosis
- Incident repair
- Incident recovery
- Normal service restoration

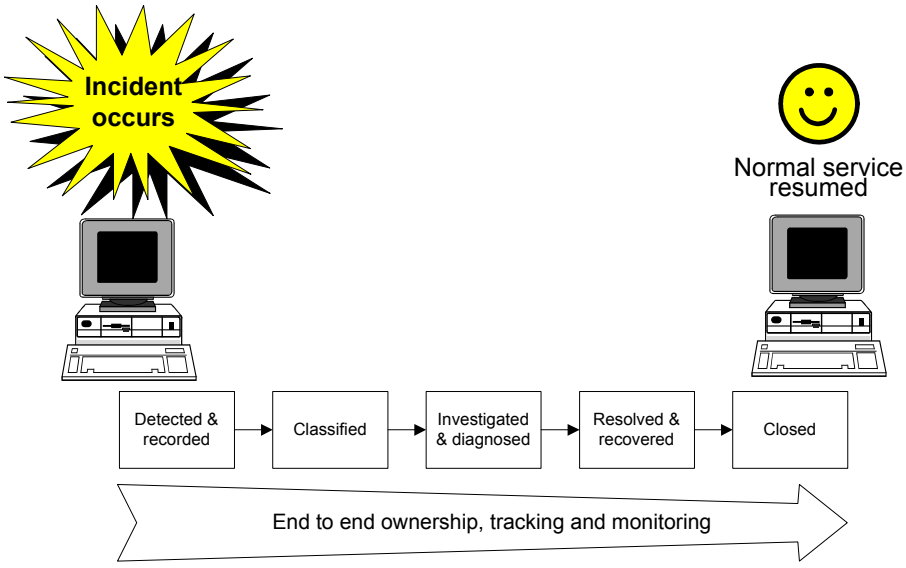


Figure 3

Incident life cycle

The time taken during each of these stages affects the overall period of downtime due to this incident and the availability of the IT service as a whole. Designing for recovery is concerned with the efficient handling of each stage in this life cycle for every IT component involved in the support of critical business functions and transactions.

Incident Life Cycle

The following figure shows the relationship between each of the incident life cycle stages and how the amount of time spent in each stage affects the overall period of downtime and the availability of the IT service as a whole.

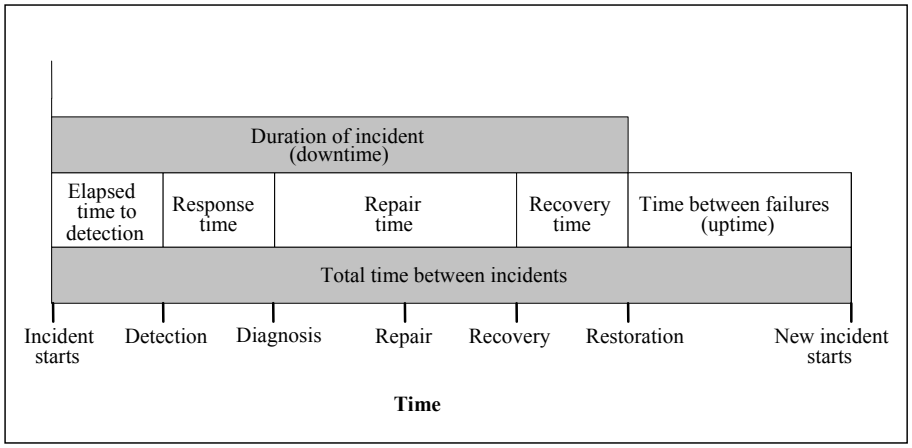


Figure 4

Incident life cycle showing time between failures

The existence of one or more countermeasures may alleviate some or all of the negative impacts of an availability incident, but the overall recovery life cycle remains predominantly the same. The repair and recovery processes may be altered as, by definition, service is still being provided by the IT infrastructure and this may restrict the nature of some of the diagnostic and recovery attempts that are allowed. In addition, reverting back from the countermeasure to normal running may involve scheduled downtime and this needs to be negotiated with the customer. In both cases, the duration of the incident is be artificially lengthened but this actually serves to maximize availability.

The role of designing for recovery is to examine each stage in an incident's life cycle and to minimize the time spent in each area. This work should be undertaken from the following two main perspectives.

Generic MOF Processes and Procedures

Clearly, the quality of the management processes within the operating and supporting phases of the MOF process model is crucial to the speed with which incidents are detected and resolved. Designing for recovery needs to ensure that these generic processes are of a sufficient level of maturity to support the needs of the IT service being implemented.

Specific Needs of Each IT Component

Continuing with the earlier focus on the key components that make up the IT service being considered, each of these components needs to be considered in detail and their requirements for optimum recovery analyzed and satisfied throughout each stage of the incident life cycle.

Following is a detailed description of each stage of the incident life cycle, including examples of relevant availability design considerations:

Incident Start

The moment that a business function or transaction becomes unavailable or degrades beyond the definition of availability agreed with the customer. This could be due to a failure of an IT component, environmental problems such as power failure, an application error, or human error such as inadvertently shutting down the wrong server.

Incident Detection

The moment that the IT organization becomes aware of the problem.

This is typically through the generation and receipt of an error message or some other audible or visual alert. Efficient event handling mechanisms need to be in place to ensure that the incident is identified as soon as possible and is not lost among other “noise” events. Pro-active threshold alerting provides advance notice of performance and capacity issues. Ideally, the IT component itself will be instrumented to advise of warning conditions.

Although inevitable with respect to certain failure types, the worst scenario from an availability perspective is to rely on the user calling in to report an outage.

Incident Diagnosis
The moment at which the true cause has been identified as opposed to any initial symptoms. Diagnosis includes the time taken to respond to the detected event and to identify appropriate resources to work on identifying the cause and to get them into a position where they can interact with the system.

Once engaged, specialists need access to a knowledge base of known problems, accurate configuration information, recent change history, appropriate diagnostics tools, and an effective escalation path and contacts list.

Incident Repair
The moment at which any underlying failure or system issue has been repaired or worked around.

Using the earlier incident examples, repair might mean the replacement of an IT component, the restoration of power, implementation of an emergency application fix, or the restart of a server. Considerations include off-hours call rotation, appropriate contracts with internal groups and external vendors, spare equipment on site, and so on.

Repair does not mean that the IT service is fully available once more or indeed that it is even back up and running.

Incident Recovery
The moment at which any recovery has been completed and the IT component is ready to resume normal processing.

For example, a replacement disk drive needs to have its data restored either from backups or from an on-line process before it can be used for production. Considerations include the provision of detailed recovery processes for IT components and the maintenance of appropriate interrelationships and dependencies.

Incident Restoration
The moment at which normal service is restored and the business function or transaction becomes fully available.

This requires synchronization with the customer and a means of communication with all users.

Designing for Customer Satisfaction During Outages

It is important to note that good customer satisfaction can still be maintained at times of failure and despite periods of unscheduled downtime. The key is to:

- Establish appropriate and realistic expectations during the requirements definition phase of the availability management life cycle.
- Articulate a clear understanding of the circumstances under which the service can be expected to fail, and how this is related to the resources being spent to protect it.

Clearly, if the IT service never approaches the levels of availability agreed with the customer, then they have a right to be dissatisfied. If the reasons for failure and the manner in which the failures are handled fall within expectations, then satisfaction is maintained.

An efficient process for handling and recovering from failures, coupled with a good clear communication path to the customer, is required. The customer needs to be kept informed at regular intervals during any recovery processing. Realistic timescales need to be given for when the service is expected to resume.

Management Processes

Many of the considerations for availability in regard to management processes are already covered by previously mentioned availability design activities. The impact on the availability of IT services of the management processes within the MOF model is significant enough to warrant separate mention.

Availability management needs to ensure that the MOF processes used for the support of critical IT services are mature enough and have the necessary people, skills, and tools to effectively undertake their respective responsibilities. The design process should look in detail at each of the management processes involved in the support of the IT service being considered.

An effective tool to help with this responsibility and also with the complete availability design process is an availability review or assessment service from an outside organization specializing in availability management, ITIL, and MOF. Such a service can help establish a baseline of maturity for any existing IT infrastructure and compare and contrast this to the needs of new or existing IT services being deployed.

Formalize Operating Level Agreements

When IT and the customer agree on a cost-effective level of availability, the agreement needs to be formalized in a document called an operating level agreement (OLA). The OLA serves as one of the building blocks for the SLA between IT and the customer. The SLA is a two-way written agreement between an IT Service Provider and the IT customer(s), defining the key service targets and responsibilities of both parties. The OLA is an agreement between internal IT service providers.

The agreed upon understanding and definition of availability, the mechanism for reporting availability back to the customer, the degree of granularity required, the format and style of presentation, and the frequency of reporting need to be formally agreed upon.

With regards to new IT services and the introduction of major change, the availability goals and metrics agreed upon with the customer must be turned into an effective set of acceptance test criteria to help prove that any new systems and support infrastructure being deployed are capable of meeting the goals and objectives that have been set for them. This work also includes acceptance tests for the role of availability management and its ability to effectively monitor and report on the new IT service.

The OLA needs to include:

- A definition of the business processing provided.
- Importance to the organization.
- The number of users.
- The business impact of downtime or unavailability.
- The cost of downtime or unavailability and how these costs change over time.
- Hours of service required.
- Critical periods of service: peaks, month-end, deadline processing and so on.
- Less critical periods of service where downtime is more tolerable.
- Scheduled downtime periods for planned maintenance and upgrades.
- Length of time downtime can be tolerated before contingency plans need to be invoked.
- How availability is to be measured, including:
- Definition of up-time.
- Definition of downtime.
- Minimum performance characteristics required.
- Minimum access points to be provided.
- How availability is to be reported to the business and how often.

This work is undertaken in close cooperation with the service level management function, as the service level manager is ultimately responsible for the negotiation and documentation of service levels with the customer.

Roles and Responsibilities

Principal roles and their associated responsibilities for availability management have been defined according to industry best practice. Organizations might need to combine some roles, depending on organizational size, organizational structure, and the underlying SLAs existing between the IT department and the business it serves.

Availability Manager

Main Responsibilities

The availability manager is responsible for managing the activities of the availability management process. This individual is responsible for ensuring that any given IT service delivers the levels of availability agreed upon with the customer and for interfacing with all other management processes in pursuit of this goal.

The availability manager:

- Ensures customer requirements are correctly translated into realistic availability goals.
- Ensures appropriate IT budgets are established for protecting the service.
- Oversees planning activities in relation to designing for availability and designing for recovery.
- Ensures that all risks to availability are identified and appropriately handled.
- Undertakes availability modeling to help select the most appropriate countermeasures, assesses the impact of future changes, and identifies potential improvements.
- Implements cost-effective countermeasures to single points of failure where possible.
- Ensures that remaining gaps are identified to the customer and ultimately handled by service continuity management when required.
- Ensures that the overall IT infrastructure is mature enough to support the availability needs.

Defines the need for and helps with the implementation of availability monitoring processes and tools.

- Ensures availability goals are reflected within appropriate service level agreements both inside and outside the company.
- Manage the day-to-day availability requirements of services.
- Collects and interprets availability metrics on behalf of the customer.
- Forecasts the impact of future availability requirements.
- Participates personally or through a delegate in the change advisory board to review the availability impact of proposed business and infrastructure changes.
- Manages a continuous availability improvement process.
- Provides consulting expertise for the review and creation of any external contracts that include availability clauses.

Relationship to Other Processes

Availability management is one of the foundational service management functions (SMFs) in the MOF optimizing quadrant. This quadrant seeks to negotiate SLAs with customers and optimize the IT infrastructure, possibly initiating requests for changes (RFCs) to the IT infrastructure. The list of SMFs in the MOF optimizing quadrant is shown below.

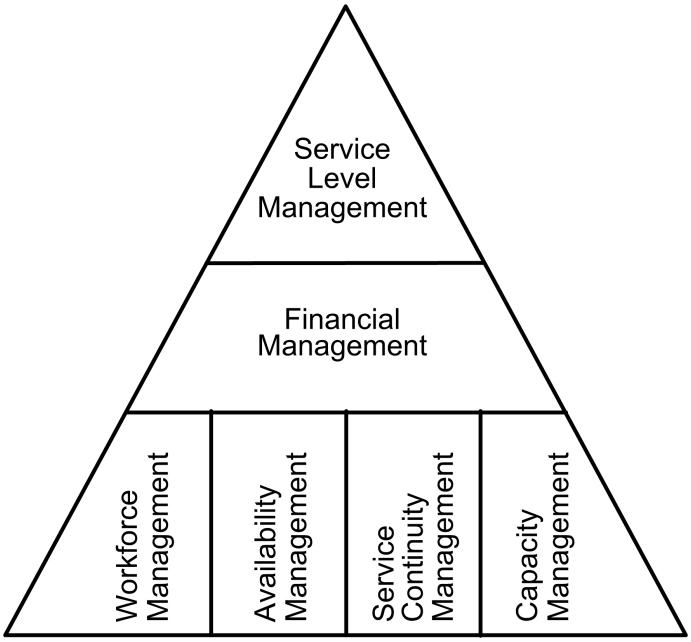


Figure 5
Microsoft Operations Framework optimizing quadrant

Service Level Management

Service level management negotiates and manages SLAs and OLAs, of which availability is a component.

From an availability perspective, service level management takes primary responsibility for interfacing with the customer and determining which IT services are most crucial to the survival of the company. Availability management draws on this prioritization work and:

- Identifies the key IT infrastructure components that support these critical services.
- Determines whether they contain any single points of failure or other risks to availability that can be cost-effectively addressed through the use of appropriate countermeasures.

Where no straight-forward countermeasures are available or where the countermeasure is prohibitively expensive or beyond

the scope of a single IT service to justify in its own right, then these availability risks are passed to service continuity management to handle.

Financial Management

Financial management acts a filter, ensuring that solutions proposed by availability management, capacity management, or service continuity management can be justified in terms of their cost to implement versus their benefit to the customer. Financial management strives to monitor, control, and, if necessary, recover costs incurred by the IT organization.

Workforce Management

Whenever a new technology is introduced into the IT environment, the people that run that technology must be properly trained and motivated. Workforce management ensures that existing personnel are trained and ready to operate a new availability solution when it is ready.

Service Continuity Management

Availability management and service continuity management are closely related as both processes strive to eliminate risks to the availability of IT services. The prime focus of availability management however, is handling the routine risks to availability that can be expected on a day-to-day basis such as the failure of a hardware component. Service continuity management caters to the more extreme and relatively rare availability risks such as fire or flood.

Capacity Management

Capacity management ensures that appropriate IT resources are available to meet customer requirements by planning for additional resources as current system resource use begins to near the point of full capacity. Availability management has a very close tie to this process, since optimal use of IT resources to meet performance levels at a justifiable cost relates to the result of effective availability management. Availability reporting and measurement highlight availability trends indicating capacity or performance issues.

Change Management

Capacity management ensures that appropriate IT resources are available to meet customer requirements by planning for additional resources as current system resource use begins to near the point of full capacity. Availability management has a very close tie to this process, since optimal use of IT resources to

meet performance levels at a justifiable cost relates well to the result of effective availability management. Availability reporting and measurement highlight availability trends indicating capacity or performance issues.

Contributors

Many of the practices that this document describes are based on years of IT implementation experience by Accenture, Avanade, Microsoft Consulting Services, Fox IT, Hewlett-Packard Company, Lucent Technologies/NetworkCare Professional Services, and Unisys Corporation.

Microsoft gratefully acknowledges the generous assistance of these organizations in providing material for this document.

Program Management Team

William Bagley, Microsoft Corporation

Jeff Yuhas, Microsoft Corporation

Lead Writers

James Westover, Hewlett Packard Corporation

Ashley Hanna, Hewlett Packard Corporation

Contributing Writers

William Bagley, Microsoft Corporation

Vicky Howells, Fox IT

Jeff Yuhas, Microsoft Corporation

Editors

Patricia Rytkonen, Volt Technical Services

Sybil Wood, Volt Technical Services