

On-Call Handbook

Best practices and tricks to get you through your on-call duties

<https://aka.ms/oncallhandbook>

On-Call Handbook**Version:** 1.002**Author:** Alaks Sevugan**Feedback:** driculture@microsoft.com**Contributors:** Sandeep Kanapala, Chad Townes, Tanner Lund, Jason Johnson**Reviewers:** Oren Rosenbloom, Zhangwei Xu, Niall Murphy, Duke Kamstra,
Matt Loflin, Surendra Shetty

Prepare to go On Call

Add IcM's number [425-421-6669](tel:425-421-6669) to your Contacts/Favorites list

Enable [Do Not Disturb](#) at night and [Allow Calls From > Favorites](#)

Install/update the [IcM Mobile App](#). Test Notifications

Ensure [your credentials](#) are current for accessing your systems and tools

Ensure you can [access the SAW machine](#): log in, update, & verify VPN access

Review the previous week's incidents. Attend your [Service Livesite Review](#)

<https://aka.ms/oncallhandbook>

Install the IcM Mobile App



aka.ms/icm.android



aka.ms/icm.ios



aka.ms/icm.intune

<https://aka.ms/icmmobile>

CONTENTS

Chapter 1: What's On-Call?	6
Introduction	6
What's On-Call?	7
Terminology	8
Chapter 2: Prepare	9
Preparing for On-Call Rotation	9
Setting up Phone Notifications.....	11
iPhone Users.....	11
Android Users.....	12
All Users	12
Setting up IcM Notifications.....	13
Business Hours	13
Non-Business Hours.....	14
Chapter 3: Act	16
Incident Workflow	16
Handling Incident Notifications	16
How to check for Outages?	19
Handling Request Assistance.....	21
Handling Incident Transfers	22
Mitigating Incidents	24
Working through a Crisis.....	25
What is a Crisis?	25
Being engaged for a Crisis	25

Getting help during a crisis.....	26
Bridge Mechanics During a Crisis impacting multiple teams.....	26
The Technical Control Bridge.....	26
The Partner Control Bridge.....	27
The Comms Control Bridge.....	27
Once you have joined the bridge, follow the best practices outlined below:	27
Postmortems.....	28
What is a Postmortem?.....	28
What Postmortem is not?	28
When are Postmortems required?	29
DRI Responsibilities.....	29
Steps to create the postmortems in IcM.....	30
Postmortem Template	32
Relaxation	35
Chapter 4: Hand-Off.....	36
Handing over On-Call responsibilities.....	36
Finish any postmortems.....	36
Relaxation	36
Chapter 5: BEST PRACTICES	37

You are On-Call?

**DON'T
PANIC**

CHAPTER 1: WHAT'S ON-CALL?

INTRODUCTION

Welcome to the on-call life! Is this your first time going on-call? Or, perhaps not your first time, but you want to see what others are doing with theirs? You're in the right place.

Just like emergency rooms require on-call schedules for doctors to handle health emergencies, **online services** need on-call schedules to efficiently respond to software and system issues that impact performance, deployment, and availability. On-call schedules ensure issues are not missed and are assigned to people with the skills to effectively evaluate and initiate a plan of corrective action.

Being on-call can be a daunting experience for any new team member. There is only one thing worse than being woken up at 3 am to discover that your systems are down — to wake up on your own at 8 am and discover that your systems were down for 5 hours and nobody got the notification or picked up the incident. While on-call shouldn't be a soul-destroying experience, some on-call rotations are more stressful than others, and there are things you can do to make your life easier and less strenuous.

This handbook is a collection of best practices and tricks to get you through your on-call duties from people who have been there.

WHAT'S ON-CALL?

Time on-call is a fact of life working in a DevOps environment, but for everyone, it's the most challenging part of the job. Working with a 24/7 platform, on-call means getting up in the middle of the night, interrupting weekend time, and putting personal life on hold. And it's stressful! It's easy to feel alone during a crisis, not wanting to bother colleagues or managers but needing help, advice, or just another set of eyes.

Being on-call means that you can be contacted at any time to investigate, mitigate and fix issues that may arise for the services you are responsible for. On-call responsibilities extend beyond normal office hours (typically 24x7) and if you are on-call, you are expected to be able to respond to issues, even at 3 am. This sounds horrible (and it sometimes can be), but this is important for our Organization and our customers.

For example, if you are on-call for a service in Azure Networking should any monitor be triggered for Sev 2 or worse issues for that service, you will receive notifications on your mobile device via email, phone call, push notification or SMS. You should acknowledge the incident first thereby taking ownership of the incident. Notification will include a link to the IcM incident which you can view in the IcM mobile app or in the Web portal to understand the details on what's wrong and how to fix it.

In C+AI, Service teams are expected to take whatever actions are necessary to mitigate the issue within **30 minutes** and return their service to a state help reduce the customer impact and build customer trust.

TERMINOLOGY

DRI	Designated Responsible Individual (or OCE)
OCE	On-Call Engineer
SAW	Secured Access Workstation. A secured laptop that is required to access production machine in Azure
IcM	One Incident Management system used at Microsoft
CEN	Classification, Escalation, and Notification matrix details how to assess the severity of an outage, when to engage specific response teams, and how frequently to send notification updates. For Azure, this is at http://aka.ms/AzureCEN
TTA	Time to Acknowledge. Time delta between when an incident is assigned to the team or request assistance was initiated and when the DRI acknowledges the incident or request. Refer to https://aka.ms/azmetrics for more details
TTJ	Time to join the bridge. This is the time delta between when DRI acknowledges the escalation and when the DRI joins the bridge. Refer to https://aka.ms/azmetrics for more details
TTE	Time to Engage. The time it takes to engage the team that mitigates impact. Refer to https://aka.ms/azmetrics for more details
TTM	Time to mitigate an incident. Time delta between when an incident first affects customers (internal or external) and when the issue has been fixed for the customers. Refer to https://aka.ms/azmetrics for more details
TSG	Troubleshooting Guides

CHAPTER 2: PREPARE

PREPARING FOR ON-CALL ROTATION

As a DRI or on-call engineer, you have a critical role to play in incident management. In the first few minutes of the incident occurring, you will need to understand the severity and customer impact of the incident. Your actions can mean the difference between an incident turning critical or being managed and resolved quickly. Preparing for your on-call rotation is a critical step for having a good week during your on-call.

On an individual level, make sure that you explain what on-call means to your friends, family members, partners, pets, etc. Make sure you make up for any missed friend/family time once your shift is over, and if you can, consider setting up a silent alarm (like a smartwatch) that can wake you up by buzzing your wrist to avoid waking up anyone around you as well. Find ways of taking care of yourself during your on-call shift and when it is over. You might want to put together an “on-call emergency survival kit” of things that help you relax: listen to a playlist of your favorite music, read a favorite book, or set aside time to play with a pet.

On an Organization level, be aware of our TTA/TTJ/TTE/TTM/TTN metrics and plan any activities accordingly. If you have 5-minute window to respond/acknowledge your notification, do so as soon as reasonably possible. Never intentionally wait until the latest allowed moment. You should follow these best practices to get ready for your On-Call rotation:

1. **Download** the [IcM Mobile App](#)¹ and **ensure that the ringer** is turned on high.
2. **Use the instructions in** [Setting up otifications](#) to set up your phone and configure your IcM notification settings.
3. **Do a** [Test Notification via ICM](#)², and **accept** the IcM On-Call Meeting Invite
4. **Ensure your credentials** are current. Confirm all permissions, passwords & expiration dates, reinstall all required tools [refer to your team's runbook] scan for changes/updates.
5. **Ensure your SAW machine** has a full charge, log in, update, & verify off of CorpNet (test VPN).
6. **Understand how C+AI handles** high severity incidents, as well as what the different roles and methods of communication. Overview of Crisis Management response can be found at <http://aka.ms/CMEPlaybook>.
7. **Be aware of your upcoming on-call** time (primary, backup) and arrange swaps, using [IcM's substitution](#) ³ feature, around travel, vacations, appointments etc.
8. **Know your escalation path:** Make sure you have **all your necessary escalation points documented**. Save copies of critical team contacts (peers, managers, & leaders).
9. **Attend your Service Livesite Review** & review previous week's incidents.
10. **Review your team's** Health Dashboards.
11. **Connect with your backup** and verify they are completing this checklist as well.
12. Other necessary best practices for preparing for your On-Call rotation:

¹ <https://icmdocs.azurewebsites.net/mobile/installation.html>

² <https://icmdocs.azurewebsites.net/administration/oncall/testMyContactInfo.html>

³ <https://icmdocs.azurewebsites.net/administration/oncall/subs.html>

- a. Prepare to be unable to complete the most trivial tasks during your "day job" as you may be tired from late night calls or being constantly distracted.
- b. Now is a good time to delegate work to colleagues.

Be prepared to be patient. Sometimes you will be called for issues outside your domain of responsibility. The person calling you is not doing this to annoy you, they need help and they are not sure what the severity of the incident is and where to route the incident to. Help them as best you can and work together to solve the issue. If this issue is not in your Service help to transfer the incident to the appropriate team. If your team notices that this is a common issue where teams are assigning the incident to you incorrectly, then work with the icmsupport@microsoft.com to get recommendations on how to resolve this challenge.

SETTING UP PHONE NOTIFICATIONS

This section covers the tips for setting up your phone and your notification profiles in IcM. Following guidance assumes that you are using an iPhone or Android cell phone and IcM.

IPHONE USERS

1. Add IcM's number **425-421-6669** to your Contacts list and add it as a Favorite.
2. In your iPhone's Settings, enable **Do Not Disturb** at night and **Allow Calls From > Favorites**. This way you can sleep soundly, knowing that if you get a phone call, it's either someone important (from your Favorites) or issues with your service.

ANDROID USERS

1. Add IcM's number **425-421-6669** to your Contacts list.
2. Starting with Android Marshmallow, you can enable **Do Not Disturb** in "Priority Only" mode to only allow notifications from contacts you specify.
3. In addition to **Do Not Disturb** mode (or as an alternative, for pre-Marshmallow phones), you can also use [Tasker](http://tasker.dinglish.net/index.html)⁴ to customize notifications: for example, when certain numbers call or text, use Tasker to *raise the volume to max* (good for heavy sleepers!). There are many ways to do this, but [here is one example](http://www.androidauthority.com/tasker-emergency-calls-399762/)⁵.

ALL USERS

1. Setup your IcM notifications to have its own unique ringtone. If you rely only on SMS or push notifications, you will constantly worry about missing them (or miss them) and your heart will race every time your phone rings. IcM is highly customizable, and a combination of email, SMS, push, and phone calls are recommended.
2. Set your IcM push sounds and ringtone to be something highly distinct from your regular ringtone and avoid commonly used or OS-native sounds. This good help avoids confusing a stranger's SMS tone in public with your own on-call notifications.
3. Make sure your phone is not silenced.

Note that your cell phone may not always work, or SMS could be delayed, or the phone call could go to voicemail directly. You should ensure good escalation policy is in place and have multiple notification mechanisms set up.

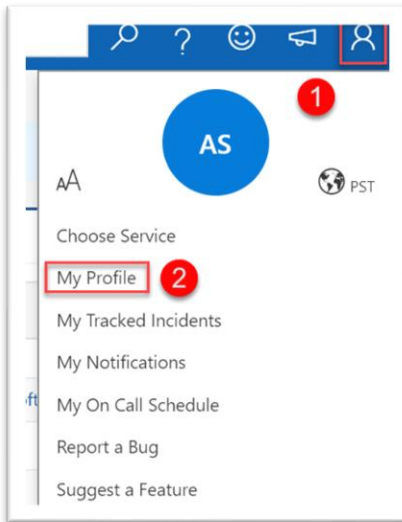
⁴ <http://tasker.dinglish.net/index.html>

⁵ <http://www.androidauthority.com/tasker-emergency-calls-399762/>

SETTING UP ICM NOTIFICATIONS

You can set up your notification profiles that are catered to your order-of-notification preference, as well as your "dexterity issues" during the wee hours.

In IcM, choose the 'My Profile' option to update your notification profile:



BUSINESS HOURS

In the 'Notification Profile' tab, update the notification preferences as suggested below, so that you are notified via multiple methods (Email, Push notifications, SMS, Mobile) in 5 minutes during business hours.

Business Hours

Set Hours

✉	Immediately	Email me at	alsevuga@microsoft.com	
💬	Immediately	SMS me at		
📱	After 2 minutes	Call my mobile at		
📱	After 2 minutes	Call my mobile at		

+ Add another action

Save

NON-BUSINESS HOURS

In the *'Notification Profile'* tab, update the notification preferences as suggested below, so that you are notified via multiple methods (Email, Push notifications, SMS, Mobile) in 5 minutes during non-business hours.

After Hours

✉	Immediately	Email me at	alsevuga@microsoft.com	
💬	Immediately	SMS me at	+14252832711	
📱	After 2 minutes	Call my mobile at	+14252832711	
📱	After 2 minutes	Call my mobile at	+14252832711	

+ Add another action

Save



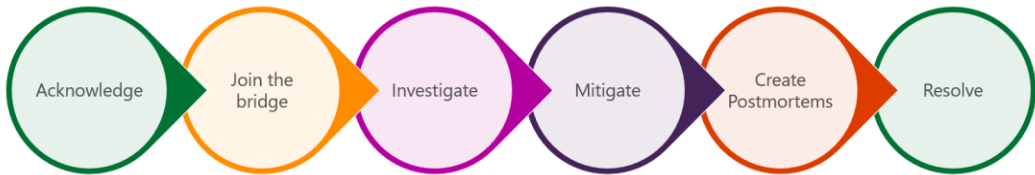
One common mistake is to have all notification types trigger at the same time. Unfortunately, what ends up happening is that in the time that it takes you to start responding to the one notification, another will trigger on top of the first. You can't acknowledge the Push Notification before the SMS comes in, and while you're trying to type the number to ACK the SMS, the call will come in.

It's for this reason that it's better to stagger notifications. Emails don't provide any special notifications (no beeps or vibrations, etc.), but it is helpful to have emails as a reference. Consider adding email as a simultaneous notification for this reason. Some incident messages may get mangled on a mobile device; having the reference in email is a good backup.

CHAPTER 3: ACT

INCIDENT WORKFLOW

Knowing exactly what your responsibilities are, can make being on-call much more painless. Below are responsibilities as they relate to each step of the incident management workflow.




Each step in the incident workflow and the expectation in each step are described in the following sections.

HANDLING INCIDENT NOTIFICATIONS

Incident notifications are sent to you based on the notification profile that you have set up in your IcM profile. This was covered in the topic: [Setting up IcM](#). Notifications in Chapter 1. When you acknowledge the incident, you are indicating that you are taking ownership of the issue.


IcM supports Voice Call, Push Notifications, SMS and emails. Each type of notification is described below.

Typical voice call message is shown here. You should press 1 or 2




Phone: IcM will call you to alert you of an incident.

- Press 1 to Acknowledge
- Press 2 to say Not Available




Push Notifications: IcM will send a push notification of the incident. Tap to launch App.

 ICM now

PRIMARY Incident Alert (Severity 2 has been raised for incident Id 50186753)

Typical IcM push notification is shown here. Tap to launch the IcM app in your phone and acknowledge the incident.

Typical IcM SMS message is shown here. You can respond to the numbers shown in the message or you can click the URL link to acknowledge the incident.



Text: IcM will text you to alert you of an incident.

- Reply back directly or click the link next to the appropriate response.

Today 11:21 AM

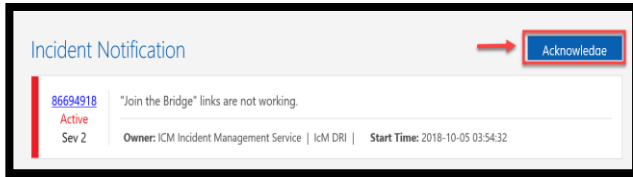
Ashley's PetShop Team needed for Sev 2 Incident 5 0 1 8 6 7 5 3 View <http://aka.ms/Ysh0db>

Reply with the number/letter, or click on the link for your chosen response.

1G: I acknowledge swnow.io/DwRGSdH

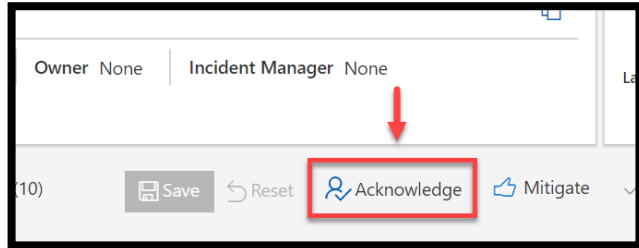
2G: I am not available swnow.io/lcEPHq

Text Message



Acknowledge button is shown in the Email notifications sent to you. You can acknowledge the incident by clicking on the 'Acknowledge' button in the email.

You can also 'Acknowledge' the incident by using the 'Acknowledge' button in the ICM web portal @ <https://aka.ms/icm>



Here are the best practices to follow when you receive an incident notification as a DRI:

1. **Acknowledge** and **act** on incidents whenever you can.
 - c. You're **NOT** expected to... **BE A HERO AND ACKNOWLEDGE ALL OF THE INCIDENTS**. Commuting and other necessary distractions are facts of life, and sometimes it's not possible to receive or act on an incident before it escalates. And that's okay — that's why ICM supports the backup on-call.

2. You should **wait for confirmation** that your acknowledgment is recorded, or your acknowledgment may be too late to cancel the next type of notification that you have specified in your notification profile. IcM will go ahead with other notifications if there is no acknowledge within the wait time for each notification type.
3. **Review the incident activity** in the Incident's '**Discussions**' field.
4. **Create a OneNote whiteboard** to track your investigation details. Details captured in the OneNote is very helpful when the incident is transferred to another team or when new team members join the incident bridge.
5. **Check whether there is an ongoing Outage** that is impacting your Service by using the Outage Dashboard. If there is an **ongoing** Outage, then join the bridge for the Outage and indicate that your service is also impacted. Once you've reported that your service is impacted, continue to execute your service's mitigation procedures. Do not wait for your dependency to mitigate their outage.

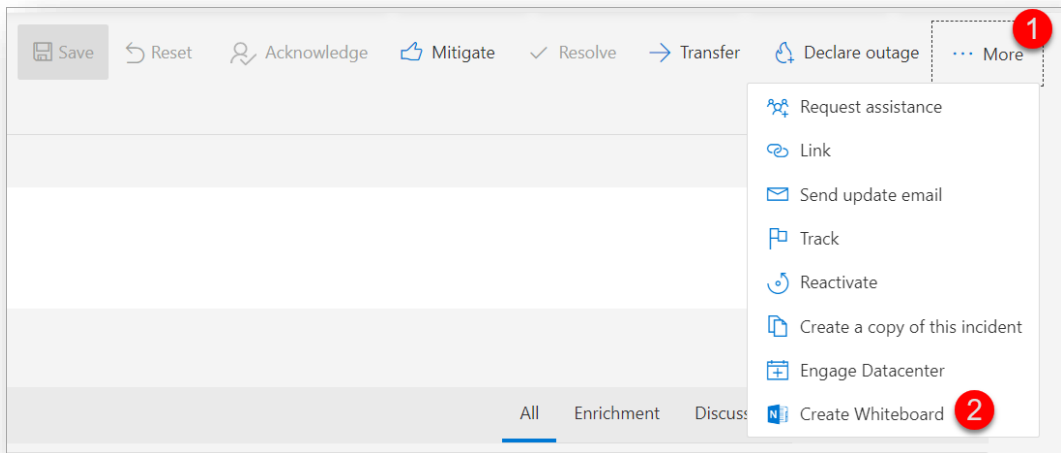
HOW TO CREATE A ONENOTE WHITEBOARD?

IcM provides the ability to create a Whiteboard in a OneNote and populate the initial content of the whiteboard using a template. Creating the Whiteboard allows the DRIs and other individuals to capture the investigation notes in one place. This helps new people joining the incident bridges and ensures that the investigation details and progress are retained when the incident is transferred from CSS (Support) teams to Service teams or when the incidents are transferred between dependent services impacted by the outages.

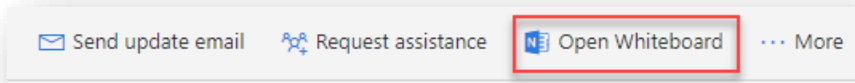
NOTE: Not creating the OneNote whiteboard leads to constant disruption and frustration for new team members joining the call or when the incident is transferred to another team.

For Incidents, you may create and/or open the whiteboard using these steps.

1. Open the incident
2. Go to the **...More** menu
3. Click the **Create Whiteboard** or **Open Whiteboard** option to open the OneNote based whiteboard.



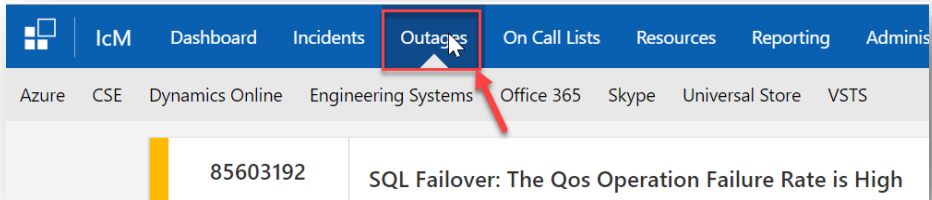
For Outages, you may create and/or open the whiteboard for an outage click the **Open whiteboard** button in the button bar



HOW TO CHECK FOR OUTAGES?

If you are looking for Outages that may be (or may have) affecting your service you can use the IcM dashboard. Outages are categorized by Microsoft brands.

1. Click on the **Outages** menu in IcM.



2. Choose the Microsoft brand you are interested in (Azure, CSE, VSTS etc).
3. Outages dashboard lists the active outages for that Microsoft brand.

Azure Outages

Declared Outages Potential Outages Interesting Incidents

Severity: 2 or Lower Time period: 4 Weeks Auto Refresh: OFF 1m 5m 10m

Status	ID	Title	Start Time/Duration	Owning Service	Location	Bridge
MITIGATED Severity 2	87670600	[RunnerHealth]StorageHealthcheck.PROD.EastUS is less than 85% healthy	10/12/2018 19:13 PDT 1h 10m	PowerApps		
MITIGATED Severity 1	87665785	BL2 Outage	10/12/2018 18:22 PDT 4h 27m	Xstore	East US	Join the Bridge

NOTE: If you are looking for an outage related to a service that isn't part of one of the pre-defined brands you can search for All outages using this link: <https://icm.ad.msft.net/imp/v3/outages/dashboard/all>.

HANDLING REQUEST ASSISTANCE

Requests are sent by teams or your team members when they need help resolving an issue. When a request assistance is made for an incident, there will be a

"Request Assistance" notification sent to the target team or person based on the severity of the incident and the user's notification profile. In some situations, the incident managers or crisis managers running an Outage can force voice notification when submitting the request to reach teams who have disabled 'Request Assistance' capability.

When a "Request Assistance" notification is received by you,

1. **Acknowledge** the Request Assistant request
2. Then, **check whether there is a bridge link** in the notification.
 - a. If a bridge link **is available** in the incident notification, then join the bridge as soon as you can help with the incident investigation.
 - b. If the bridge link is **NOT available**, then acknowledge the notification for high severity incidents and for low severity incidents, provide appropriate SLA for the response.

HANDLING INCIDENT TRANSFERS

Incidents are typically transferred to you when the Owning team has triaged the incident and determined that the incident needs to be mitigated by your service. Since high severity incidents must be mitigated in less than 30 mins, start triaging the incident to find the root cause as soon as possible. If the incident is determined to be caused by a different team, then transfer the incident to the right team using IcM's [transfer assistance](#)⁶ experience as detailed below:

1. Open the incident in IcM.
2. Click on the '*Transfer*' menu option.

⁶ <https://icmdocs.azurewebsites.net/workflows/Incidents/TransferAssistance.html>

Hit Count
0

Last Correlated
None

Impact Duration
0m

Start Time
10/07/2018 15:48 PDT

[Add New Bridge](#)

[Acknowledge](#)
[Mitigate](#)
[Resolve](#)
→ Transfer
[Declare outage](#)
[More](#)

3. Select from the 'Recommended Teams' (option 2) or 'Previous owners' (option 3) or 'Search for Service/Team' (option 1).

Transfer Ownership ⓘ

Important: If this incident has other incidents correlated to it, this operation will affect all correlated incidents as well.

Service/Team(s) ⓘ 1

Scope by: Service Contact All

Scope the search by service...

Search services or teams

☀ Recommended Owners (2) ⓘ
🕒 Previous Owners (1) ⓘ 3

2	Service	Team	Match Score ▼
<input type="radio"/>	CE Security Engineering	SonarSRE	99.95% <div style="width: 99.95%;"></div>
<input type="radio"/>	IAM Services	ID Operations	70% <div style="width: 70%;"></div>

Reason *

Reason

[Cancel](#)
[Transfer](#)

4. Provide the `Reason` for the transfer.
5. Click on 'Transfer' to transfer the incident.

MITIGATING INCIDENTS

Every incident must be prioritized. To prioritize an incident, start by assessing its impact on the business based on the number of subscriptions and customers that are impacted. As a DRI, you don't have to necessarily fix a problem on your own. Ask for help from your team as soon as possible.

The C+AI organization has a goal of mitigating the incident in less than 30mins and your role as the DRI is critical in achieving this goal. Use the following guidelines when an incident is assigned to you.

2. **Ask yourself, does the incident and your initial investigation indicate a general problem** or an issue with a specific service that you should investigate?
 - a. If it does not look like a problem you are the expert in, then IcM's [request assistance](#)⁷ feature to ask for help from the team member who owns the component or service.
 - b. If the issue is not in your service, then transfer the incident to the appropriate team using IcM's [Transfer Assistance](#)⁸ feature to transfer the incident to the right team.
 - i. Make sure you thoroughly add notes in IcM's discussion field, on what steps you've taken, your conclusions, and WHY you are transferring to this specific team. In these moments where minutes matter, we do not want the next person to spend time re-doing work or figuring out why they were called.
3. **Determine the urgency of the issue:**

⁷ <https://icmdocs.azurewebsites.net/workflows/Incidents/Requesting%20Assistance.html>

⁸ <https://icmdocs.azurewebsites.net/workflows/Incidents/TransferAssistance.html>

- a. Is it something that should be worked on right now and perhaps declared as an outage?
- b. Is it some tactical work that doesn't have to happen during the night? For example, if there's a performance issue limited to a very small set of requests and the trend is not indicating impending doom, suppress the monitors using IcM's [suppression rule](#)⁹ until for a more suitable time, such as normal working hours and get back to fixing it then.
- c. If you decide the work doesn't need to happen immediately, create a repair item to make sure to change your monitors so that you don't get woken by similar incidents in the future.

WORKING THROUGH A CRISIS

WHAT IS A CRISIS?

In the context of Azure Services, Outages at the Sev 1+ level are broadly defined in the [Azure CEN](#) ¹⁰ as multi-service/multi-region incidents. These types of incidents are classified as Crises. Crises in Azure are run by Azure Crisis Managers. Examples of Crisis include the multi-service/multi-region scenarios, loss of physical infrastructure, potential Data Loss, major security issues, and issues with the potential for brand-level impact. Azure Services are expected to update the Azure CEN with their Major and Minor scenarios. To update the scenarios, please send an email to waom@microsoft.com.

BEING ENGAGED FOR A CRISIS

Crisis Managers will invite the teams to the Outage either by creating incidents for each team or by sending a request assistance to the on-call DRIs needed for

⁹ <https://icmdocs.azurewebsites.net/administration/rules/alert%20suppression%20rules.html>

¹⁰ <https://aka.ms/azurecen>

restoration. When you receive the request, you must join the active engineering bridge as soon as possible. If you are not able to join, then the backup DRI or someone from your team must join. You can go to <https://aka.ms/dritraining> to learn about the DRI best practices described here.

GETTING HELP DURING A CRISIS

A service team with a Sev 1 + incident as defined in the [Azure CEN¹¹](#) (examples above) may connect with Crisis Management through IcM using a Request Assistance to Windows Azure Operations Center / Incident Manager. This will result in a Crisis Manager engaging on your bridge to understand impact and risk.

BRIDGE MECHANICS DURING A CRISIS IMPACTING MULTIPLE TEAMS

There are 3 major bridges that are hosted during a multi-service Crisis:

THE TECHNICAL CONTROL BRIDGE

The Technical Control Bridge, or TCB, is the focal point for technical restoration of an issue. This involves the service teams necessary for restoration of services that are impacted up the stack. Services close to the impact chain, or required as next-step recovery services are also asked to be on the bridge to test lower-level mitigation and begin their own steps to recovery should their recovery be gated on restoration below the stack. Service teams that are not necessary for technical restoration of the core will be asked to join the Partner Control Bridge. TCB is managed by Sr.IM from Crisis Management team to define and drive the workstreams necessary to mitigate the impact. Or in few cases the IM team or Executive IM of the service that is causing impact manages this bridge.

¹¹ <https://aka.ms/azurecen>

THE PARTNER CONTROL BRIDGE

The Partner Control Bridge, or PCB, is the rally point for all service teams that are not key to restoration of the platform during a crisis. This will include both First, and Third-party workloads. Additionally, field teams may join the PCB. Azure's Customer Experience Team leads the PCB, which is a Chat Only bridge (i.e. Voice is not enabled). When the TCB begins testing restoration, it will request validation from teams on the PCB.

THE COMMS CONTROL BRIDGE

The Comms Control Bridge, or CCB, is a chat only bridge comprised of Crisis Management, CXP Communications, and Marketing/PR DRIs. CCB is managed by Comms Managers team and is used to ensure that the appropriate communications are being crafted to support a world-class customer experience during a crisis. It is rare that an Azure Service team will be asked to join this bridge.

ONCE YOU HAVE JOINED THE BRIDGE, FOLLOW THE BEST PRACTICES OUTLINED BELOW:

1. **Always type the following key data into the chat:** The start time, end time, workstreams, changes in progress, etc.
2. **Do not “ghost” the bridge:** If you need to step away from the bridge at any time for any reason, declare in the chat how long you will be away.
3. **Do not acknowledge an escalation if you cannot join the bridge:** Always inform your backup using the “I am not available” feature in IcM's notification if you are commuting or unable to perform your Live Site duties in full.

4. **Be active on the bridge:** Type if you are not comfortable speaking on the bridge. Let the Crisis Manager know if you are using a translator.
5. **Do not have side discussions or offline investigations with your team:** All workstreams and discussions should take place on the bridge unless instructed otherwise.

POSTMORTEMS

WHAT IS A POSTMORTEM?

Postmortem process seeks to improve the overall quality of our services, tools and processes by documenting the timeline to identify root causes and address them through trackable repair items. As our systems scale and become more complex, failure is inevitable, assessment and remediation are more involved and time-consuming, and it becomes increasingly painful to repeat recurring mistakes.

Postmortems help us:

1. To find the root cause of the issue via diving deep into the sequence of events that happened leading to the incident.
2. To address the root cause (code, design, process, tools etc) of the issue via trackable and deliverable repair items.
3. To stop the re-occurrence of the problem.
4. To analyze the impact of the problem on our business and our customers.
5. To capture the learnings from our analysis to share within the team and across other teams in C+AI.

WHAT POSTMORTEM IS NOT?

1. Postmortem is not a process for finding whom to blame for the incident.
2. “Removing blame from a postmortem gives people the confidence to escalate issues without fear. An atmosphere of blame risks creating a culture in which incidents and issues are swept under the rug, leading to

greater risk for the organization.” from Site Reliability Engineering book by Niall Murphy.

3. Postmortem is not a process for giving punishment to employees after the occurrence of a bad event.

NOTE: Regardless of what we discover, we understand and truly believe that everyone did the best job they could, given what they knew at the time, their skills and abilities, the resources available, and the situation at hand.

WHEN ARE POSTMORTEMS REQUIRED?

Postmortems are required for all **Sev 0 – 2** incidents that are declared as an Outage. Teams can choose to create the postmortems for other incident severities to share the learnings within their team and to improve the quality of their service.

DRI RESPONSIBILITIES

During a Live Site event, your team is 100% focused on restoring the service as soon as possible. You cannot, and should not, be wasting time and mental energy on thinking about how to do something more optimally, nor performing a deep dive on figuring out the root cause of an outage. That’s why post-mortems are essential, providing a peacetime opportunity to reflect once the issue is no longer impacting customers. The postmortem process drives focus, instills a culture of learning, and identifies opportunities for improvement that otherwise would be completely lost.

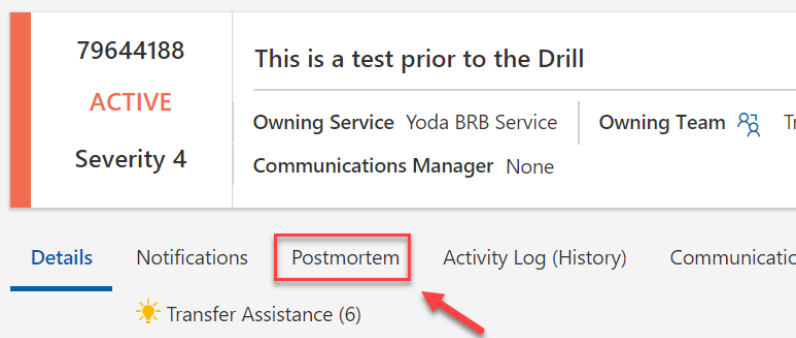
As a DRI, here are your responsibilities for the postmortems:

1. You are responsible for creating the postmortems and assigning it to the Service owners within your team who will be responsible for authoring the postmortems.

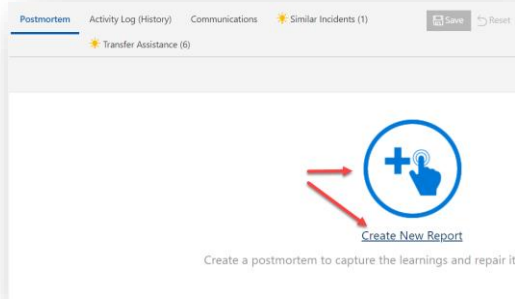
2. You must add notes in the incident to help the postmortem author to understand what happened during the live site event.
3. You must update the timelines in the incident to reflect the sequence of events.
4. If root causes are available, then update the incident to include the root causes.
5. DRIs must be present during live site reviews and (where possible) actual incident reviews/debriefings.

STEPS TO CREATE THE POSTMORTEMS IN ICM.

1. Click on the **Postmortem** tab in the incident details



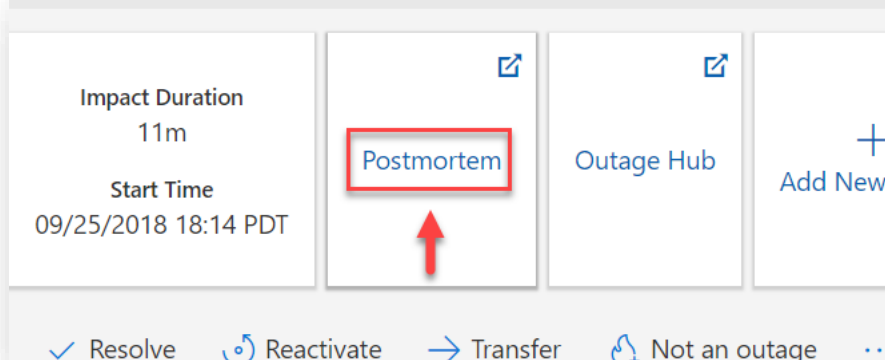
2. Click on **Create New Report** button in the Postmortem view. This creates a new postmortem for the incident.



- By default, the postmortem is assigned to you. Change the **Owner** field to the right owner and save the postmortem. No other fields are required.

The screenshot shows the 'Internal Postmortem' form with the 'Title & Ownership' tab selected. The form has seven steps: 1. Title & Ownership, 2. Impact, 3. Timeline, 4. Root Cause, 5. Detection & Mitigation, 6. Repair Items, and 7. Additional Details. The 'Title' field contains the text 'This is a test prior to the Drill'. The 'Severity' dropdown is set to '4'. The 'Owning Team' dropdown is set to 'Triage'. The 'Owner' dropdown is set to 'Alaks Sevugan (alsevuga)' and is highlighted with a red circle and a red arrow. The 'Incident Manager' and 'Communications Manager' dropdowns are empty. A red arrow points from the 'Owner' field to the 'Save and Continue' button at the bottom right.

- Postmortem is now created and can be accessed from the top card



POSTMORTEM TEMPLATE

Postmortems must be blameless, be focused on learnings and continuous improvements. Postmortems will be read by multiple audiences, your team members, your leadership, other services impacted by this incident, senior leadership at Live site meetings. So, make sure that the postmortem are simple to understand for all readers.

In general, an effective postmortem must tell a story and should include the following sections:

- 1. Summary**

- a. A 1-2 paragraph executive summary of the incident, covering: who, what, where and why. Also, include a short (1-2 sentence) introduction that helps orient readers to your service or function.

- 2. Metrics / Graphs**

- a. Includes the graphs from Geneva Monitoring or AppInsights or Azure Monitor that demonstrates the impact of the event. For example. Fatal errors, latency, API throttling etc.

3. Customer Impact

- a. A 1-2 paragraph summary of the customer impact during the Live Site. Include details of the number of subscriptions and the number of customers impacted. Identify the Pulse customers impacted by the incident. If you do not have a metric for your service, add a repair item to be addressed.

4. Incident Response Analysis

Add a section that includes details about the detection and diagnosis of the event.

- a. How was the Live Site detected (e.g. our monitors, another team's monitors, manual)?
- b. How could time to detection be improved? how would you have cut the time in half?
- c. How did you reach the point where you knew how to mitigate the impact?
- d. How could time to mitigation be improved?

5. Root Causes

- a. Choose 1 or more root causes that caused the Live Site incident.
- b. For changes triggered by a deployment or a change, complete the following:
 - i. Is the change automated? If not, why?
 - ii. If the change was automated, should this have been caught and rolled back in testing?
- c. Did you have an existing backlog item that would've prevented or greatly reduced the impact of this Live Site?
 - i. If yes, why was the item not completed prior to the Live Site?
- d. Is there an existing Service 360 violation that would have prevented this Live Site if addressed?

- i. If yes, why was it not addressed?
- ii. If no, is it possible to programmatically audit for the vulnerability or failure mode that you experienced?

6. Timeline

Include all the time events in the timeline (impact start, a time when the issue was detected, mitigation time etc) along with a short description.

7. '5 Whys' section to deep dive into the issue to identify the root cause and repair items

- a. Start with the problem. Keep asking why until you get to the root cause and repair items. Add an example to help the users think through the problem:
- b. Let's say your website is down. Here are the 5 whys you could ask:
 - i. **Why was the website down?**
The CPU utilization on all our front-end servers went to 100%.
 - ii. **Why did the CPU usage spike?**
A new bit of code contained an infinite loop!
 - iii. **Why did that code get written?**
So-and-so made a mistake since he is a new developer.
 - iv. **Why did his mistake get checked in?**
We did not have sufficient quality gates to prevent this check-in without unit tests.
 - v. **Why didn't we have sufficient quality gates?**
We did not have check-in verifications to verify that there was not enough code coverage for check-ins.

8. Lessons Learned

- a. What are the key takeaways from this incident?
- b. How can you prevent this from happening again in the future?
- c. What processes broken down?

- d. Is there tooling or process improvements required to prevent this from happening in the future?

9. Repair Items

- a. Use your learnings from the 5 whys and determine a list of repair items required for fixing this problem.
- b. If an issue keeps happening but turns out not be customer impacting, consider improving it as a longer-term task. This can include incidents for disks that fill up, logs that should be rotated, noisy monitors, etc.
- c. All repair items must be created in VSTS, directly from IcM.
- d. Repair items that should be addressed in the next **2 weeks** should be marked as a short-term repair item.
- e. If the information is difficult to find, create repair items to capture the information in your team's TSGs so that you can constantly refactor and improve your TSGs or documentation.

RELAXATION

On-call can be STRESSFUL. Whether you're getting hammered with incidents/calls or you only got one (but it was at 3 am) getting into a relaxed and sleepy frame of mind can be difficult.

Make sure to sleep as much as you can during the On-Call week.

If you are going to be late to arriving in the office because you worked on an issue late into the night, make sure to let your backup DRI know.

CHAPTER 4: HAND-OFF

HANDING OVER ON-CALL RESPONSIBILITIES

When your on-call “shift” ends let the next on-call know about issues that have not been resolved yet and other experiences of note. Provide a quick summary to the next on-call team member about any issues that may come up during their shift. This can be a written report via email or in a team’s OneNote or a verbal summary

Ensure to make a positive hand-off to the next On-Call individual and make them aware of the ‘Active Incidents’. In addition, provide a list of postmortems that you are working on. Finally, follow the procedures outlined by your team’s post-rotation guidelines.

FINISH ANY POSTMORTEMS

You own completing the postmortems for any incidents that occurred during your on-call rotation. Make sure you complete them while events are fresh in your mind.

RELAXATION

Once your on-call rotation is done, it’s crucial that you take time to recharge—to completely unplug yourself from your day-to-day routine and allow your body and mind to truly rest. Most DRIs don’t realize they are burning out until they completely burn out.

CHAPTER 5: BEST PRACTICES

Use the following best practices to optimize your on-call rotations:

PRIORITIZE REPEAT INCIDENTS TO BE FIXED TO AVOID DRI BURNOUT: A sustainable on-call is only possible if the developers building the system place importance on designing *reliability* into a system - reliability cannot be addressed in an on-call shift. On-Call can “suck” if any of the following issues are not prioritized - noisy alerts, alerts that aren’t actionable, alerts missing crucial bits of information, outdated runbooks or non-existent runbooks, not resolving the repair items, lack of logs to diagnose the problem etc

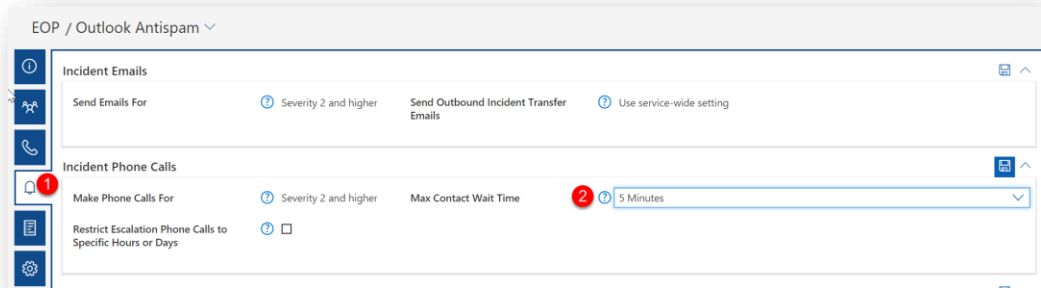
ALWAYS HAVE A BACKUP IN YOUR SCHEDULE Yes, this means a minimum of two people are on-call at the same time; A backup shift should generally come directly after a primary shift. It gives chance for the previous primary to pass on the additional context which may have come up during their shift. It also helps to prevent people from sitting on issues with the intent of letting the next shift fix it.

SHADOWING FOR NEW MEMBERS: New members of the team should shadow their on-call rotation during the first few weeks. They should get all incidents and should follow along with what you are doing.

GET YOUR TEAM INVOLVED: Have an escalation policy of at least 2 levels deep setup. This should include the *Incident Manager* Team and the *Executive Incident Manager* team, so they are aware of what is going on. This should hopefully never happen, but when it does, it’s useful to be able to get a hold of the next available person.

MAX CONTACT WAIT TIME is not more than 5 mins. We recommend you set your contact wait timeout to 5 minutes. This should be plenty of time for someone to acknowledge the incident if they’re able to. If they’re not able to within 5 minutes,

then they're probably not in a good position to respond to the incident anyway. This can be configured in ICM team's "Max contact wait time?" field.



IF YOU ARE MAKING A CHANGE THAT IMPACTS THE SCHEDULE, let others know since many people plan around the on-call schedule well in advance.

SUPPORT EACH OTHER: When doing activities that might generate plenty of “calls”, such as maintenance, it is courteous to “take the calls” yourself while performing the actions that will call the DRIs. Take the On-Call schedule for this time period by notifying them and scheduling an On-Call substitution in ICM, for the duration.

Prepare to go On Call

Add IcM's number [425-421-6669](tel:425-421-6669) to your Contacts/Favorites list

Enable [Do Not Disturb](#) at night and [Allow Calls From > Favorites](#)

Install/update the [IcM Mobile App](#). Test Notifications

Ensure [your credentials](#) are current for accessing your systems and tools

Ensure you can [access the SAW machine](#): log in, update, & verify VPN access

Review the previous week's incidents. Attend your [Service Livesite Review](#)

<https://aka.ms/oncallhandbook>

Install the IcM Mobile App



aka.ms/icm.android



aka.ms/icm.ios



aka.ms/icm.intune

<https://aka.ms/icmmobile>