Pexip Infinity and Amazon Web Services

Deployment Guide

Software Version 31

Document Version 31.a

March 2023

]pexip[

Contents

Introduction ... **4**

Deployment guidelines ... **5**

Deployment options ... 5

Limitations ... 6

Recommended instance types and call capacity guidelines ... 6

IP addressing ... 6

Assumptions and prerequisites ... 7

Configuring AWS security groups ... **8**

Inbound rules ... 8

Outbound rules ... 8

Inter-node communication requirements for multiple VPCs ... 8

Deploying a Management Node in AWS ... **10**

Task summary ... 10

Task breakdown ... 10

Initial platform configuration — AWS ... **14**

Accessing the Pexip Infinity Administrator interface ... 14

Configuring the Pexip Infinity platform ... 14

Next steps ... 15

Deploying a Conferencing Node in AWS ... **16**

Task summary ... 16

Deploying the VM instance in AWS ... 16

Generating, downloading and deploying the configuration file ... 19

Assigning a persistent public IP address ... 21

Configuring dynamic bursting to the AWS cloud ... **22**

Configuring your system for dynamic bursting to AWS ... 22

Firewall addresses/ports required for access to the AWS APIs for cloud bursting ... 22

Setting up your bursting nodes in AWS and enabling bursting in Pexip Infinity ... 22

Configuring an AWS user and policy for controlling overflow nodes ... 23

Configuring the bursting threshold ... 24

Manually starting an overflow node ... 25

Converting between overflow and "always on" AWS Conferencing Nodes ... 25

Managing AWS instances ... **27**

Temporarily removing (stopping) a Conferencing Node instance ... 27

Reinstating (restarting) a stopped Conferencing Node instance ... 27

Permanently removing a Conferencing Node instance ... 28

© 2023 Pexip AS

Adding ENA support to an instance 28

Viewing cloud bursting status **29**

Viewing current status 29

Viewing historic events 29

Introduction

The Amazon Elastic Compute Cloud (Amazon EC2) service provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using AWS eliminates your need to invest in hardware up front, so you can deploy Pexip Infinity even faster.

You can use AWS to launch as many or as few virtual servers as you need, and use those virtual servers to host a Pexip Infinity Management Node and as many Conferencing Nodes as required for your Pexip Infinity platform.

AWS enables you to scale up or down to handle changes in requirements or spikes in conferencing requirements. You can also use the AWS APIs and the Pexip Infinity management API to monitor usage and bring up / tear down Conferencing Nodes as required to meet conferencing demand, or allow Pexip Infinity to handle this automatically for you via its dynamic bursting capabilities.

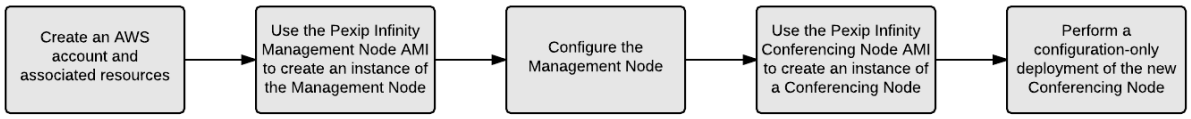
Pexip publishes Amazon Machine Images (AMIs) for the Pexip Infinity Management Node and Conferencing Nodes. These AMIs may be used to launch instances of each node type as required.

© 2023 Pexip AS

Deployment guidelines

This section summarizes the AWS deployment options and limitations, and provides guidance on our recommended AWS instance types, security groups and IP addressing options.

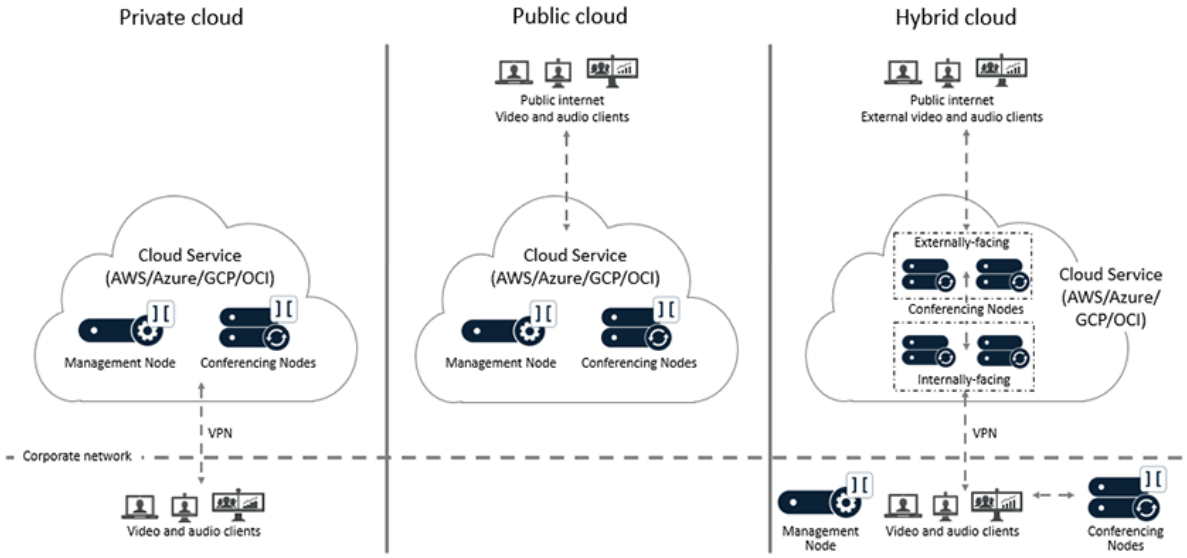
This flowchart provides an overview of the basic steps involved in deploying the Pexip Infinity platform on AWS:



Deployment options

There are three main deployment options for your Pexip Infinity platform when using the AWS cloud:

- **Private cloud**: all nodes are deployed within an AWS Virtual Private Cloud (VPC). Private addressing is used for all nodes and connectivity is achieved by configuring a VPN tunnel between the corporate network and the AWS VPC. As all nodes are private, this is equivalent to an on-premises deployment which is only available to users internal to the organization.
- **Public cloud**: all nodes are deployed within the AWS VPC. All nodes have a private address but, in addition, public IP addresses are allocated to each node. The node's private addresses are only used for inter-node communications. Each node's public address is then configured on the relevant node as a static NAT address. Access to the nodes is permitted from the public internet, or a restricted subset of networks, as required. Any systems or endpoints that will send signaling and media traffic to those Pexip Infinity nodes must send that traffic to the public address of those nodes. If you have internal systems or endpoints communicating with those nodes, you must ensure that your local network allows such routing.
- **Hybrid cloud**: the Management Node, and optionally some Conferencing Nodes, are deployed in the corporate network. A VPN tunnel is created between the corporate network and the AWS VPC. Additional Conferencing Nodes are deployed in the AWS VPC and are managed from the on-premises Management Node. The AWS-hosted Conferencing Nodes can be either internally-facing, privately-addressed (private cloud) nodes; or externally-facing, publicly-addressed (public cloud) nodes; or a combination of private and public nodes (where the private nodes are in a different Pexip Infinity system location to the public nodes). You may also want to consider dynamic bursting, where the AWS-hosted Conferencing Nodes are only started up and used when you have reached capacity on your on-premises nodes.



All of the Pexip nodes that you deploy in the cloud are completely dedicated to running the Pexip Infinity platform— you maintain full data ownership and control of those nodes.

Limitations

The following limitations currently apply:

- Pexip Infinity node instances that are hosted on AWS can be deployed in one or many regions within AWS. However, if you deploy nodes across multiple AWS regions, it is your responsibility to ensure that there is a routable network between the AWS data centers, so that inter-node communication between the Management Node and all of its associated Conferencing Nodes can succeed. Pexip is unable to provide support in setting this AWS network up.

  Each AWS region contains multiple Availability Zones. A Pexip Infinity system location is equivalent to an AWS Availability Zone.

  Note that service providers may deploy multiple independent Pexip Infinity platforms in any AWS location (subject to your licensing agreement).
- SSH access to AWS-hosted Pexip Infinity nodes requires key-based authentication. (Password-based authentication is considered insufficiently secure for use in the AWS environment and is not permitted.) An SSH key pair must be set up within the AWS account used to launch the Pexip Infinity instances and must be assigned to each instance at launch time. You can create key pairs within AWS via the EC2 Dashboard Key Pairs option, or use third-party tools such as PuTTYgen to generate a key pair and then import the public key into AWS.

  Note that:
  - Pexip Infinity node instances only support a single SSH key pair.
  - If you are using a Linux or Mac SSH client to access your instance you must use the **chmod** command to make sure that your private key file on your local client (SSH private keys are never uploaded) is not publicly viewable. For example, if the name of your private key file is my-key-pair.pem, use the following command: `chmod 400 /path/my-key-pair.pem`

  See http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html for more information about creating a key pair.
- We do not support AWS deployments in China.

Recommended instance types and call capacity guidelines

AWS instances come in many different sizes. In general, Pexip Infinity Conferencing Nodes should be considered compute intensive and Management Nodes reflect a more general-purpose workload. Our Server design recommendations also apply to cloud-based deployments.

For deployments of up to 30 Conferencing Nodes, we recommend using:

- **Management Node**: an **m5.xlarge** instance.
- **Transcoding Conferencing Nodes**: a **c5.2xlarge** instance.
- **Proxying Edge Nodes**: either **c5.xlarge** or **c5.2xlarge**.

This should provide capacity for approximately 17 HD / 39 SD / 324 audio-only calls per Transcoding Conferencing Node.

Larger instance types may also be used for a Transcoding Conferencing Node, but the call capacity does not increase linearly so these may not represent the best value. However, we recommend that you do not use c5 or c4 instances with 36 vCPU or higher.

Note that you can switch between c4 and c5 AWS instance families for existing VMs. To do this you **must** power down the VM, change its family, and then power the VM on again.

ℹ We recommend that the Management Node and each Conferencing Node instance is run as a dedicated instance (tenancy).

For more information see:

- Amazon EC2 instance types: https://aws.amazon.com/ec2/instance-types/
- Amazon EC2 dedicated instances pricing guidelines: https://aws.amazon.com/ec2/pricing/dedicated-instances/

IP addressing

Within a VPC, an instance's private IP addresses can initially be allocated dynamically (using DHCP) or statically. However, after the private IP address has been assigned to the instance it remains fixed and associated with that instance until the instance is terminated. The allocated IP address is displayed in the AWS management console.

Public IP addresses may be associated with an instance dynamically (at launch/start time) or statically through use of an Elastic IP. Dynamic public IP addresses do not remain associated with an instance if it is stopped — and thus it will receive a new public IP address when it is next started.

Pexip Infinity nodes must always be configured with the private IP address associated with its instance, as it is used for all internal communication between nodes. To associate an instance's public IP address with the node, configure that public IP address as the node's **Static NAT address** (via **Platform > Conferencing Nodes**).

Assumptions and prerequisites

The deployment instructions assume that within AWS you have already:

- signed up for AWS and created a user account, administrator groups etc
- created a Virtual Private Cloud network and subnet
- configured a VPN tunnel from the corporate/management network to the VPC
- created or imported an SSH key pair to associate with your VPC instances
- created a security group (see Configuring AWS security groups for port requirements)
- decided in which AWS region to deploy your Pexip Infinity platform (these guidelines assume that all Pexip Infinity node instances that are hosted on AWS are deployed in the same AWS region).

For more information on setting up your AWS environment, see http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/get-set-up-for-amazon-ec2.html.

Configuring AWS security groups

Access to AWS instances is restricted by the AWS firewall. This may be configured by associating an instance with an AWS security group that specifies the permitted inbound and outbound traffic/ports from the group.

A minimal AWS security group that permits access to a public cloud style Pexip Infinity deployment would look similar to this:

Inbound rules

Type	Protocol	Port range	Source
SSH	TCP	22	<management station IP address/subnet>
HTTPS	TCP	443	0.0.0.0/0
Custom TCP Rule	TCP	1720	0.0.0.0/0
Custom TCP Rule	TCP	5060	0.0.0.0/0
Custom TCP Rule	TCP	5061	0.0.0.0/0
Custom TCP Rule	TCP	8443	<management station IP address/subnet>
Custom TCP Rule	TCP	33000-49999	0.0.0.0/0
Custom UDP Rule *	UDP	5060	0.0.0.0/0
Custom UDP Rule	UDP	40000-49999	0.0.0.0/0
Custom UDP Rule	UDP	500	<sg-12345678>
Custom UDP Rule	UDP	1719	0.0.0.0/0
Custom Protocol	ESP (50)	All	<sg-12345678>
All ICMP	ICMP	All	<management station IP address/subnet>

* Only required if you intend to enable SIP over UDP.

Outbound rules

Type	Protocol	Port range	Source
All traffic	All	All	0.0.0.0/0

Where **0.0.0.0/0** implies any source / destination, **<management station IP address/subnet>** should be restricted to a single IP address or subnet for SSH access only, and **<sg-12345678>** is the identity of this security group (and thus permits traffic from other AWS instances — the Management Node and Conferencing Nodes — associated with the same security group).

A single security group can be applied to the Management Node and all Conferencing Nodes. However, if you want to apply further restrictions to your Management Node (for example, to exclude the TCP/UDP signaling and media ports), then you can configure additional security groups and use them as appropriate for each AWS instance.

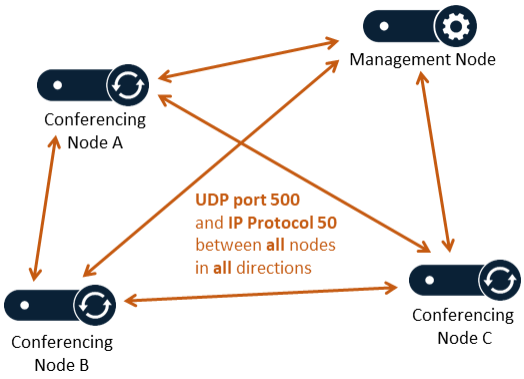
Remember that the Management Node and all Conferencing Nodes must be able to communicate with each other. If your instances only have private addresses, ensure that the necessary external systems such as NTP and DNS servers are routable from those nodes.

For further information on the ports and protocols specified here, see Pexip Infinity port usage guide.

Inter-node communication requirements for multiple VPCs

In a basic deployment, your Pexip Infinity platform will be deployed within a single VPC.

In larger deployments you may choose to deploy your Conferencing Nodes across multiple VPCs — in which case there must be a directly routable path (no NAT) between all nodes that allows **UDP port 500 (IKE)**, and **IP Protocol 50 (IPsec ESP)** to pass between all nodes in both directions.



Deploying a Management Node in AWS

As with all Pexip Infinity deployments, you must first deploy the Management Node before deploying any Conferencing Nodes. In a hybrid cloud deployment the Management Node may be deployed in the corporate network or in the AWS VPC. This section describes how to deploy the Management Node in AWS.

Task summary

Deploying a Management Node in AWS consists of the following steps:

1. In the AWS management console, pick the desired AWS region and use the launch wizard to create an instance of the Management Node.
2. Search the Community AMIs section for the relevant Pexip Infinity Management Node AMI.
3. Ensure that the instance is associated with a suitable security group, and that an SSH key pair has been associated with the instance.
4. After the instance has booted, SSH into it and set the administrator password. This will then terminate the SSH session.
5. SSH in to the Management Node again and complete the Pexip Infinity installation wizard as for an on-premises deployment.

These steps are described below in more detail.

Task breakdown

1. In the AWS management console, ensure that you have selected the AWS region in which you intend to deploy the Management Node and all of its associated Conferencing Nodes.
2. From the EC2 dashboard, select **Images > AMIs**.
3. Choose an Amazon Machine Image (AMI):
   a. Select **Public images**.
   b. Filter on "Owner : 686087431763" to see all of the Pexip images.
   c. Select the row for **Pexip Infinity Management Node <version> build <build_number>** where **<version>** is the software version you want to install. (You may also want to filter on the version number to refine the list of images.)
   d. Select **Launch**.

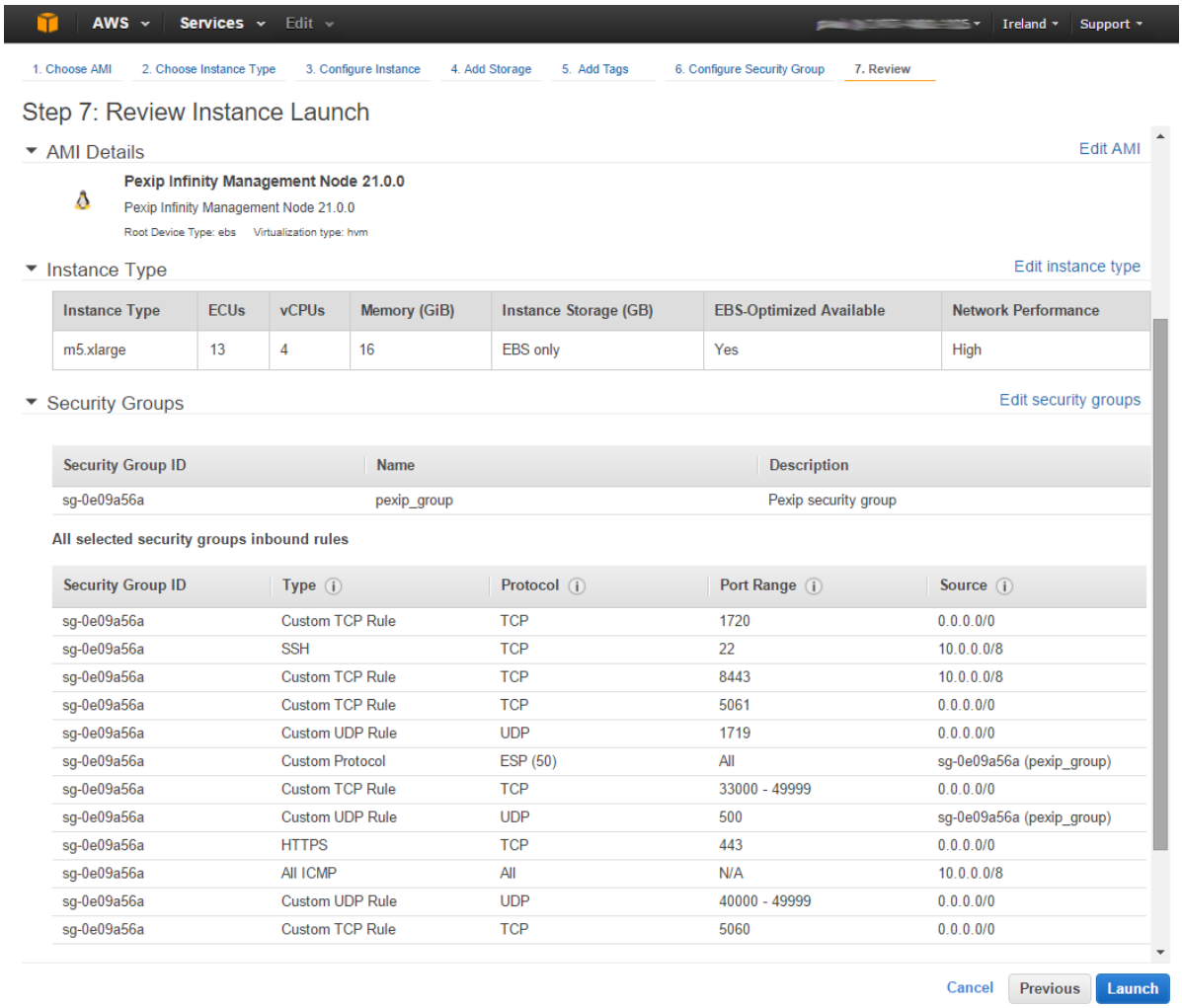
      This launches a wizard in which you will select and configure your image.
4. Complete Step 2: Choose an Instance Type:
   a. For deployments of up to 30 Conferencing Nodes, we recommend using an **m5.xlarge** instance type for the Management Node.
   b. Select **Next: Configure Instance Details**.
5. Complete Step 3: Configure Instance Details:

a. Complete the following fields (leave all other settings as default):

Number of instances	1
Subnet	Use default subnet.
Auto-assign Public IP	Enable or disable this option according to whether you want the node to be reachable from a public IP address. Your subnet may be configured so that instances in that subnet are assigned a public IP address by default. Note that the Management Node only needs to be publicly accessible if you want to perform system administration tasks from clients located in the public internet.
Primary IP	Either leave as ***Auto-assign*** or, if required, specify your desired IP address. (AWS reserves the first four IP addresses and the last one IP address of every subnet for IP networking purposes.)
Tenancy	Select ***Dedicated***.

b. Select **Next: Add Storage**.

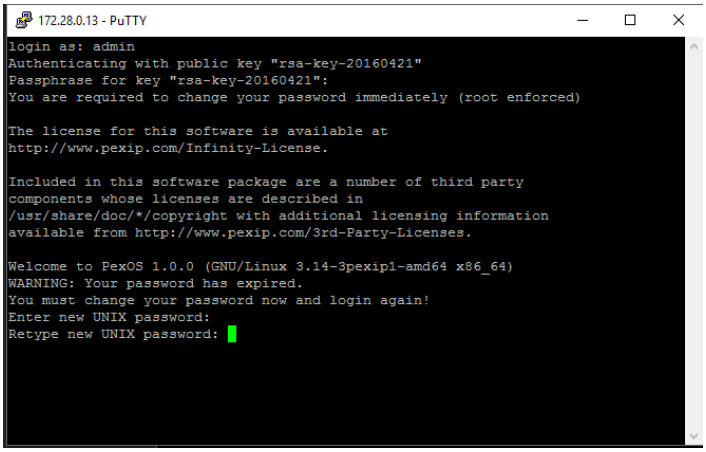
6. Complete Step 4: Add Storage:
   a. Accept the default settings (the Pexip AMI sets these defaults appropriately for a Management Node).
   b. Select **Next: Add Tags**.
7. Complete Step 5: Add Tags:
   a. You can optionally add tags to your instance, if you want to categorize your AWS resources.
   b. Select **Next: Configure Security Group**.
8. Complete Step 6: Configure Security Group:
   a. Select and assign your security group to your Management Node instance.
   b. Select **Review and Launch**.
9. Complete Step 7: Review Instance Launch:
   a. This step summarizes the configuration details for your instance.
   You may receive a warning that your security group is open to the world. This is to be expected if you are deploying a public or hybrid VPC that is intended to be accessible to publicly-located clients.



b. Select **Launch**.

10. You are now asked to select an existing key pair or create a new key pair:
    a. Select the key pair that you want to associate with this instance, and acknowledge that you have the private key file.
    You will need to supply the private key when you subsequently SSH into this instance.
    b. Select **Launch instances**.
    The **Launch Status** screen is displayed.
11. Select **View Instances** to see all of your configured instances and ensure that your Instance State is ***running***.
    The status screen also indicates the private IP address, and public IP address if appropriate, of the instance.
12. Connect over SSH into the Management Node instance to complete the installation of Pexip Infinity.
    Use an SSH client to access the Management Node by its private IP address, supplying your private key file as appropriate.
13. Follow the login process in the SSH session:
    a. At the login prompt, enter the username ***admin***.
    b. Supply the key passphrase, if requested.
    c. At the "Enter new UNIX password:" prompt, enter your desired password, and then when prompted, enter the password again.

    This will then log you out and terminate your SSH session.



14. Reconnect over SSH into the Management Node instance and continue the installation process:
    a. Log in again as ***admin***.
    You are presented with another login prompt:
    `[sudo] password for admin:`
    b. Enter the UNIX password you just created.
    The Pexip installation wizard will begin after a short delay.
    c. Complete the installation wizard to apply basic configuration to the Management Node:

IP address Network mask Gateway	Accept the defaults for the **IP address**, **Network mask** and **Gateway** settings.
Hostname Domain suffix	Enter your required **Hostname** and **Domain suffix** for the Management Node.
DNS servers	Configure one or more **DNS servers**. You must override the default values if it is a private deployment.
NTP servers	Configure one or more **NTP servers**. You must override the default values if it is a private deployment.
Web administration username Password	Set the **Web administration username** and **password**.
Enable incident reporting	Select whether or not to **Enable incident reporting**.
Send deployment and usage statistics to Pexip	Select whether or not to **Send deployment and usage statistics to Pexip**.

    ℹ The DNS and NTP servers at the default addresses are only accessible if your instance has a public IP address. The installation wizard will fail if the NTP server address cannot be resolved and reached.

    After successfully completing the wizard, the SSH connection will be lost as the Management Node reboots.
15. After a few minutes you will be able to use the Pexip Infinity Administrator interface to access and configure the Management Node (remember to use https to connect to the node if you have only configured https access rules in your security group). You can configure your Pexip Infinity platform licenses, VMRs, aliases, locations etc. as described in Initial platform configuration — AWS before you go on to add Conferencing Nodes.

Initial platform configuration — AWS

After you have run the installation wizard, you must perform some preliminary configuration of the Pexip Infinity platform before you can deploy a Conferencing Node.

This section lists the configuration required, and provides a summary of each step with a link to further information.

All configuration should be done using the Pexip Infinity Administrator interface.

No changes should be made to any Pexip VM via the terminal interface (other than as described when running the initial Pexip installation wizard) unless directed to do so by Pexip support. This includes (but is not limited to) changes to the time zone, changes to IP tables, configuration of Ethernet interfaces, or the installation of any third-party code/applications.

Accessing the Pexip Infinity Administrator interface

The Pexip Infinity Administrator interface is hosted on the Management Node. To access this:

1. Open a web browser and type in the IP address or DNS name that you assigned to the Management Node using the installation wizard (you may need to wait a minute or so after installation is complete before you can access the Administrator interface).
2. Until you have uploaded appropriate TLS certificates to the Management Node, your browser may present you with a warning that the website's security certificate is not trusted. You should proceed, but upload appropriate TLS certificates to the Management Node (and Conferencing Nodes, when they have been created) as soon as possible.

   The **Pexip Infinity Conferencing Platform** login page will appear.
3. Log in using the web administration username and password you set using the installation wizard.

You are now ready to begin configuring the Pexip Infinity platform and deploying Conferencing Nodes.

As a first step, we strongly recommend that you configure at least 2 additional NTP servers or NTP server pools to ensure that log entries from all nodes are properly synchronized.

It may take some time for any configuration changes to take effect across the Conferencing Nodes. In typical deployments, configuration replication is performed approximately once per minute. However, in very large deployments (more than 60 Conferencing Nodes), configuration replication intervals are extended, and it may take longer for configuration changes to be applied to all Conferencing Nodes (the administrator log shows when each node has been updated).

Brief details of how to perform the initial configuration are given below. For complete information on how to configure your Pexip Infinity solution, see the Pexip Infinity technical documentation website at docs.pexip.com.

Configuring the Pexip Infinity platform

This table lists the Pexip Infinity platform configuration steps that are required before you can deploy Conferencing Nodes and make calls.

Configuration step	Purpose
1. Enable DNS **(System > DNS Servers)**	Pexip Infinity uses DNS to resolve the hostnames of external system components including NTP servers, syslog servers, SNMP servers and web proxies. It is also used for call routing purposes — SIP proxies, gatekeepers, external call control and conferencing systems and so on. The address of at least one DNS server must be added to your system. You will already have configured at least one DNS server when running the install wizard, but you can now change it or add more DNS servers.

Configuration step	Purpose
2. Enable NTP (System > NTP Servers)	Pexip Infinity uses NTP servers to obtain accurate system time. This is necessary to ensure correct operation, including configuration replication and log timestamps. We strongly recommend that you configure at least three distinct NTP servers or NTP server pools on all your host servers and the Management Node itself. This ensures that log entries from all nodes are properly synchronized. You will already have configured at least one NTP server when running the install wizard, but you can now change it or add more NTP servers.
3. Add licenses (Platform > Licenses)	You must install a system license with sufficient concurrent call capacity for your environment before you can place calls to Pexip Infinity services.
4. Add a system location (Platform > Locations)	These are labels that allow you to group together Conferencing Nodes that are in the same datacenter. You must have at least one location configured before you can deploy a Conferencing Node.
5. Upload TLS certificates (Certificates > TLS Certificates)	You must install TLS certificates on the Management Node and — when you deploy them — each Conferencing Node. TLS certificates are used by these systems to verify their identity to clients connecting to them. All nodes are deployed with self-signed certificates, but we strongly recommend they are replaced with ones signed by either an external CA or a trusted internal CA.
6. Add Virtual Meeting Rooms (Services > Virtual Meeting Rooms)	Conferences take place in Virtual Meeting Rooms and Virtual Auditoriums. VMR configuration includes any PINs required to access the conference. You must deploy at least one Conferencing Node before you can call into a conference.
7. Add an alias for the Virtual Meeting Room (done while adding the Virtual Meeting Room)	A Virtual Meeting Room or Virtual Auditorium can have more than one alias. Conference participants can access a Virtual Meeting Room or Virtual Auditorium by dialing any one of its aliases.

Next steps

You are now ready to deploy a Conferencing Node — see Deploying a Conferencing Node in AWS for more information.

Deploying a Conferencing Node in AWS

After deploying the Management Node and completing the initial platform configuration you can deploy one or more Conferencing Nodes in AWS to provide conferencing capacity.

Task summary

Deploying a Conferencing Node in AWS consists of the following steps:

1. Deploying a new VM instance in AWS:
   a. In the AWS management console, select the same AWS region in which the Management Node is deployed and use the launch wizard to create an instance of a Conferencing Node.
   b. Search the Community AMIs section for the relevant Pexip Infinity Conferencing Node AMI.
   c. Ensure that the instance is run as a dedicated instance (tenancy), is associated with a suitable security group, and that an SSH key pair has been associated with the instance.
2. Configuring the VM with the details of the specific Conferencing Node being deployed, using a file generated from the Pexip Infinity Management Node.
   a. After the instance has booted, perform a configuration-only deployment on the Management Node to inform it of the new Conferencing Node.
   b. Upload the resulting XML document to the new Conferencing Node.
   c. Configure the Conferencing Node's static NAT address, if you have a assigned a public IP address to the instance.

These steps are described below in more detail.

Deploying the VM instance in AWS

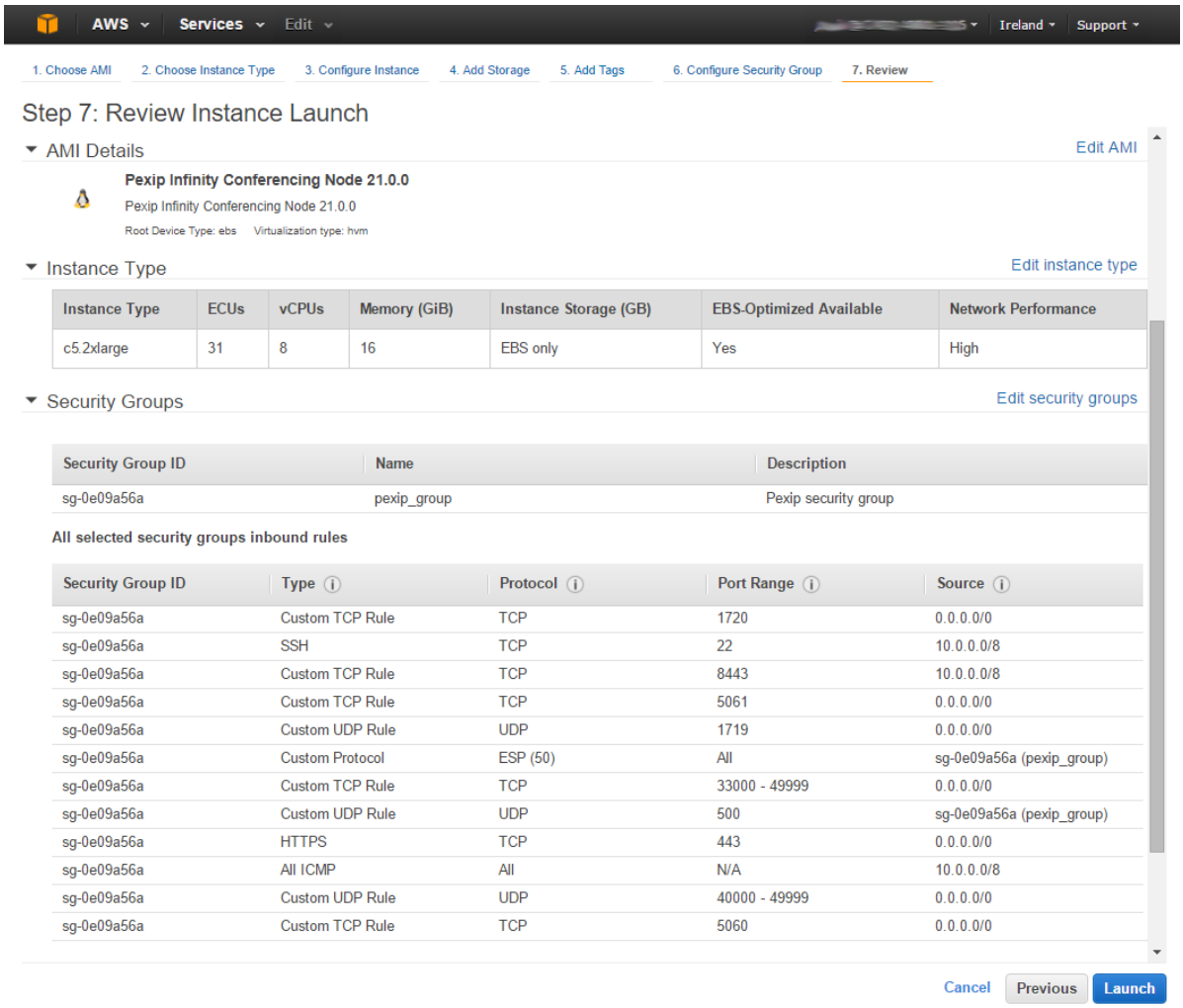
1. In the AWS management console, ensure that you have selected the same AWS region in which the Management Node is deployed.
2. From the EC2 dashboard, select **Images > AMIs**.
3. Choose an Amazon Machine Image (AMI):
   a. Select **Public images**.
   b. Filter on "Owner : 686087431763" to see all of the Pexip images.
   c. Select the row for **Pexip Infinity Conferencing Node <version> build <build_number>** where **<version>** is the software version you want to install. (You may also want to filter on the version number to refine the list of images.)
   d. Select **Launch**.
      This launches a wizard in which you will select and configure your image.
4. Complete Step 2: Choose an Instance Type:
   a. We recommend using a **c5.2xlarge** instance type for a Transcoding Conferencing Node.
      See Recommended instance types and call capacity guidelines for more information.
   b. Select **Next: Configure Instance Details**.
5. Complete Step 3: Configure Instance Details:

a. Complete the following fields (leave all other settings as default):

Number of instances	1
Subnet	Use default subnet.
Auto-assign Public IP	Enable or disable this option according to whether you want the node to be reachable from a public IP address. You must assign a static public/external IP address to the Conferencing Node if you want that node to be able to host conferences that are accessible from devices in the public internet. Your subnet may be configured so that instances in that subnet are assigned a public IP address by default. If you want to assign a persistent public IP address (an Elastic IP Address) you can do this after the instance has been launched.
Primary IP	Either leave as ***Auto-assign*** or, if required, specify your desired IP address. (AWS reserves the first four IP addresses and the last one IP address of every subnet for IP networking purposes.)
Tenancy	Select ***Dedicated***.

b. Select **Next: Add Storage**.

6. Complete Step 4: Add Storage:
   a. Accept the default settings (the Pexip AMI sets these defaults appropriately for a Conferencing Node).
   b. Select **Next: Add Tags**.
7. Complete Step 5: Add Tags:
   a. You can optionally add tags to your instance, if you want to categorize your AWS resources.
   b. Select **Next: Configure Security Group**.
8. Complete Step 6: Configure Security Group:
   a. Select and assign your security group to your Conferencing Node instance.
   b. Select **Review and Launch**.
9. Complete Step 7: Review Instance Launch:
   a. This step summarizes the configuration details for your instance.
   You may receive a warning that your security group is open to the world. This is to be expected if you are deploying a public or hybrid Conferencing Node that is intended to be accessible to publicly-located clients.



b. Select **Launch**.

10. You are now asked to select an existing key pair or create a new key pair:
    a. Select the key pair that you want to associate with this instance, and acknowledge that you have the private key file.
    (Note that you will not be required to SSH into Conferencing Node instances.)
    b. Select **Launch instances**.
    The **Launch Status** screen is displayed.
11. Select **View Instances** to see all of your configured instances and ensure that your Instance State is ***running***.
    The status screen also indicates the private IP address, and public IP address if appropriate, of the instance.
12. Make a note of the Private IP address that has been assigned to the new Conferencing Node.
13. Perform a configuration-only deployment of the new Conferencing Node as described below.

Generating, downloading and deploying the configuration file

1. From the Pexip Infinity Administrator interface, go to **Platform > Conferencing Nodes** and select **Add Conferencing Node**.
2. You are now asked to provide the network configuration to be applied to the Conferencing Node, by completing the following fields:

Option	Description
Name	Enter the name to use when referring to this Conferencing Node in the Pexip Infinity Administrator interface.
Description	An optional field where you can provide more information about the Conferencing Node.
Role	This determines the Conferencing Node's role: ◦ ***Proxying Edge Node***: a Proxying Edge Node handles all media and signaling connections with an endpoint or external device, but does not host any conferences — instead it forwards the media on to a Transcoding Conferencing Node for processing. ◦ ***Transcoding Conferencing Node***: a Transcoding Conferencing Node handles all the media processing, protocol interworking, mixing and so on that is required in hosting Pexip Infinity calls and conferences. When combined with Proxying Edge Nodes, a transcoding node typically only processes the media forwarded on to it by those proxying nodes and has no direct connection with endpoints or external devices. However, a transcoding node can still receive and process the signaling and media directly from an endpoint or external device if required.
Hostname Domain	Enter the hostname and domain to assign to this Conferencing Node. Each Conferencing Node and Management Node must have a unique hostname. The Hostname and Domain together make up the Conferencing Node's DNS name or FQDN. We recommend that you assign valid DNS names to all your Conferencing Nodes.
IPv4 address	Enter the IP address to assign to this Conferencing Node when it is created. This should be the Private IP address that AWS has assigned to the new Conferencing Node.
Network mask	Enter the IP network mask to assign to this Conferencing Node. The netmask depends upon the subnet selected for the instance. The default AWS subnet has a /20 prefix size which is a network mask of 255.255.240.0.
Gateway IPv4 address	Enter the IP address of the default gateway to assign to this Conferencing Node. This is the first usable address in the subnet selected for the instance (e.g. 172.31.0.1 for a 172.31.0.0/20 subnet).
Secondary interface IPv4 address	Leave this option blank as dual network interfaces are not supported on Conferencing Nodes deployed in public cloud services.
Secondary interface network mask	Leave this option blank as dual network interfaces are not supported on Conferencing Nodes deployed in public cloud services.
System location	Select the physical location of this Conferencing Node. A system location should not contain a mixture of proxying nodes and transcoding nodes. If the system location does not already exist, you can create a new one here by clicking + to the right of the field. This will open up a new window showing the **Add System Location** page.
SIP TLS FQDN	A unique identity for this Conferencing Node, used in signaling SIP TLS Contact addresses.
TLS certificate	The TLS certificate to use on this node. This must be a certificate that contains the above **SIP TLS FQDN**. Each certificate is shown in the format **<subject name> (<issuer>)**.
IPv6 address	The IPv6 address for this Conferencing Node. Each Conferencing Node must have a unique IPv6 address.

Option	Description
Gateway IPv6 address	The IPv6 address of the default gateway. If this is left blank, the Conferencing Node listens for IPv6 Router Advertisements to obtain a gateway address.
IPv4 static NAT address	Configure the Conferencing Node's static NAT address, if you have a assigned a public/external IP address to the instance. Enter the public IP address allocated by AWS. See Assigning a persistent public IP address below if you want the node to have a persistent public IP address (an Elastic IP address).
Static routes	From the list of **Available Static routes**, select the routes to assign to the node, and then use the right arrow to move the selected routes into the **Chosen Static routes** list.
Enable distributed database	This should usually be enabled (checked) for all Conferencing Nodes that are expected to be "always on", and disabled (unchecked) for nodes that are expected to only be powered on some of the time (e.g. nodes that are likely to only be operational during peak times).
Enable SSH	Determines whether this node can be accessed over SSH. ***Use Global SSH setting***: SSH access to this node is determined by the global **Enable SSH** setting (**Platform > Global Settings > Connectivity > Enable SSH**). ***Off***: this node cannot be accessed over SSH, regardless of the global **Enable SSH** setting. ***On***: this node can be accessed over SSH, regardless of the global **Enable SSH** setting. Default: ***Use Global SSH setting***.

3. Select **Save**.
4. You are now asked to complete the following fields:

Option	Description
Deployment type	Select ***Generic (configuration-only)***.
SSH password	Enter the password to use when logging in to this Conferencing Node's Linux operating system over SSH. The username is always ***admin***. Logging in to the operating system is required when changing passwords or for diagnostic purposes only, and should generally be done under the guidance of your Pexip authorized support representative. In particular, do not change any configuration using SSH — all changes should be made using the Pexip Infinity Administrator interface.

5. Select **Download**.
   A message appears at the top of the page: "The Conferencing Node image will download shortly or click on the following link".
   After a short while, a file with the name **pexip-<hostname>.<domain>.xml** is generated and downloaded.
   Note that the generated file is only available for your current session so you should download it immediately.
6. Browse to **https://<conferencing-node-ip>:8443/** and use the form provided to upload the configuration file to the Conferencing Node VM.
   If you cannot access the Conferencing Node, check that you have allowed the appropriate source addresses in your security group inbound rules for management traffic. In public deployments and where there is no virtual private network, you need to use the public address of the node.
   The Conferencing Node will apply the configuration and reboot. After rebooting, it will connect to the Management Node in the usual way.
   You can close the browser window used to upload the file.

After deploying a new Conferencing Node, it takes approximately 5 minutes before the node is available for conference hosting and for its status to be updated on the Management Node. Until it becomes available, the Management Node reports the status of the Conferencing Node as having a last contacted and last updated date of "Never". "Connectivity lost between nodes" alarms relating to that node may also appear temporarily.

Assigning a persistent public IP address

If you want the node to have a persistent public IP address you can assign an Elastic IP address to the Conferencing Node.

Note that the public IP address assigned when the instance was launched (if **Auto-assign Public IP** was selected), will always be available and will not change while the instance remains running. A new (different) public IP address is only assigned if the instance is stopped and restarted.

1. Assign an Elastic IP address to the instance via the **Elastic IPs** option in the Amazon VPC console.
2. Update the Conferencing Node's static NAT address:
   a. Log in to the Pexip Infinity Administrator interface on the Management Node.
   b. Go to **Platform > Conferencing Nodes** and select the Conferencing Node.
   c. Configure the **Static NAT address** as the instance's Elastic IP address as appropriate.

Configuring dynamic bursting to the AWS cloud

Pexip Infinity deployments can burst into the Amazon Web Services (AWS) cloud when primary conferencing capabilities are reaching their capacity limits, thus providing additional temporary Conferencing Node resources.

This provides the ability to dynamically expand conferencing capacity whenever scheduled or unplanned usage requires it. The AWS cloud Conferencing Nodes instances are only started up when required and are automatically stopped again when capacity demand normalizes, ensuring that AWS costs are minimized.

For complete information about dynamic bursting, see Dynamic bursting to a cloud service.

Configuring your system for dynamic bursting to AWS

These instructions assume that you already have a working Pexip Infinity platform, including one or more primary (always on) Conferencing Nodes in one or more system locations. These existing Conferencing Nodes can be deployed using whichever platform or hypervisor you prefer.

Firewall addresses/ports required for access to the AWS APIs for cloud bursting

Access to the AWS APIs for cloud bursting is only required from the Management Node.

The Management Node always connects to destination port 443 over HTTPS.

DNS is used to resolve the AWS API addresses. Currently, Pexip Infinity uses the "Amazon Elastic Compute Cloud (Amazon EC2)" DNS FQDNs listed at http://docs.aws.amazon.com/general/latest/gr/rande.html#ec2_region but this may change in the future. An exception is if you are using GovCloud, where **ec2.us-gov-west-1.amazonaws.com** is used instead.

Setting up your bursting nodes in AWS and enabling bursting in Pexip Infinity

You must deploy in AWS the Conferencing Nodes that you want to use for dynamic bursting, and then configure the Pexip Infinity location containing those nodes as the overflow destination for the locations that contain your primary (always on) Conferencing Nodes:

1. In Pexip Infinity, configure a new system location for media overflow e.g. "AWS burst", that will contain your bursting Conferencing Nodes.

   (Note that system locations are not explicitly configured as "primary" or "overflow" locations. Pexip Infinity automatically detects the purpose of the location according to whether it contains Conferencing Nodes that may be used for dynamic bursting.)
2. In AWS, set up a user and associated access policy that the Pexip Infinity Management Node will use to log in to AWS to start and stop the node instances.

   See Configuring an AWS user and policy for controlling overflow nodes for more information.
3. Deploy in AWS the Conferencing Nodes that you want to use for dynamic bursting. Deploy these nodes in the same manner as you would for "always on" usage (see Deploying a Conferencing Node in AWS), except:
   a. Apply to each cloud VM node instance to be used for conference bursting a tag with a **Key** of **pexip-cloud** and an associated **Value** set to the **Tag value** that is shown in the **Cloud Bursting** section on the **Platform > Global Settings** page (the **Tag value** is the hostname of your Management Node).

      This tag indicates which VM nodes will be started and shut down dynamically by your Pexip system, and relates to the access policy document configured in the previous step.
   b. When adding the Conferencing Node within Pexip Infinity:
      i. Assign the Conferencing Node to the overflow system location (e.g. "AWS burst").
      ii. Disable (uncheck) the **Enable distributed database** setting (this setting should be disabled for any nodes that are not expected to always be available).
   c. After the Conferencing Node has successfully deployed, manually stop the node instance on AWS.
4. In Pexip Infinity, go to **Platform > Global Settings > Cloud Bursting**, enable cloud bursting and then configure your bursting threshold, minimum lifetime and other appropriate settings for AWS:

Option	Description
Enable bursting to the cloud	Select this option to instruct Pexip Infinity to monitor the system locations and start up / shut down overflow Conferencing Nodes hosted in your cloud service when in need of extra capacity.
Bursting threshold	The bursting threshold controls when your dynamic overflow nodes in your cloud service are automatically started up so that they can provide additional conferencing capacity. When the number of additional HD calls that can still be hosted in a location reaches or drops below the threshold, it triggers Pexip Infinity into starting up an overflow node in the overflow location. See Configuring the bursting threshold for more information.
Tag name and Tag value	These read-only fields indicate the tag name (always **pexip-cloud**) and associated tag value (the hostname of your Management Node) that you must assign to each of your cloud VM node instances that are to be used for dynamic bursting. In some circumstances the **Tag value** may auto populate as "unknown" instead of the hostname, in which case you should also use "unknown" on your cloud VM node instances.
Minimum lifetime	An overflow cloud bursting node is automatically stopped when it becomes idle (no longer hosting any conferences). However, you can configure the **Minimum lifetime** for which the bursting node is kept powered on. By default this is set to 50 minutes, which means that a node is never stopped until it has been running for at least 50 minutes. If your service provider charges by the hour, it is more efficient to leave a node running for 50 minutes — even if it is never used — as that capacity can remain on immediate standby for no extra cost. If your service provider charges by the minute you may want to reduce the **Minimum lifetime**.
Cloud provider	Select ***AWS***.
AWS access key ID and AWS secret access key	Set these to the **Access Key ID** and the **Secret Access Key** respectively of the **User Security Credentials** for the user you set up in the AWS dashboard within **Identity And Access Management** in step 2 above.

5. Go to **Platform > Locations** and configure the system locations that contain your "always on" Conferencing Nodes (the nodes/locations that initially receive calls) so that they will overflow to your new "AWS burst" location when necessary.

   When configuring your principal "always on" locations, you should normally set the **Primary overflow location** to point at the bursting location containing your overflow nodes, and the **Secondary overflow location** should normally only point at an always-on location.

   Nodes in a bursting location are only automatically started up if that location is configured as a **Primary overflow location** of an always-on location that has reached its capacity threshold. This means that if a bursting location is configured as a **Secondary overflow location** of an always-on location, then those nodes can only be used as overflow nodes if they are already up and running (i.e. they have already been triggered into starting up by another location that is using them as its **Primary overflow location**, or you have used some other external process to start them up manually).

   We recommend that you do not mix your "always on" Conferencing Nodes and your bursting nodes in the same system location.

Configuring an AWS user and policy for controlling overflow nodes

Within AWS you must set up a user and an access policy to be used by Pexip Infinity to start up and shut down the Conferencing Node overflow instances:

1. From the AWS management console, select **IAM** (Identity and Access Management).
2. Set up the access policy for the overflow node instances:
   a. Select **Policies** and then **Create policy**.
   b. Select the JSON tab and copy/paste the following text:

```
{
   "Version": "2012-10-17",
```

```
    "Statement": [
        {
            "Action": [
                "ec2:Describe*"
            ],
            "Effect": "Allow",
            "Resource": "*"
        },
        {
            "Action": [
                "ec2:StartInstances",
                "ec2:StopInstances"
            ],
            "Effect": "Allow",
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "ec2:ResourceTag/pexip-cloud": "<bursting-tag-value>"
                }
            }
        }
    ]
}
```

   c. Replace **<bursting-tag-value>** with the **Tag value** that is shown in the **Cloud Bursting** section on the **Platform > Global Settings** page. (This is the only element of the policy JSON that you need to change.)
   d. Select **Review policy**.
   e. Enter a policy **Name** and **Description**.
   f. Select **Create policy**.
3. Create a new user on behalf of the Pexip platform and associate it with the access policy.
   a. Select **Users** and then **Add user**.
   b. Enter a **User name** such as "**pexip**" and select an **Access type** of ***Programmatic access***.
   c. Select **Next: Permissions**.
   d. On the **Set Permissions** page, select the **Attach existing policies directly** tab.
   e. Use the **Filter policies** field to search for the policy you have just created above, and then select the checkbox next to that policy.
   f. Select **Next: Tags**, where you can optionally add some tags.
   g. Select **Next: Review** where you can review the user details and its associated permissions/policy.
   h. Select **Create user**.
   i. Either download the user credentials or show and make a note of the **Access key ID** and the **Secret access key** — you will enter these values into the **Global Settings** page in the Pexip Infinity Administrator interface.

      (You must copy or download these key values when you create the user; you will not be able to access them again later.)
   j. Finally, select **Close**.

ℹ This policy only allows the "pexip" user i.e. the Pexip Infinity platform, to retrieve a list of instances and to start and stop existing instances that you have tagged as **pexip-cloud**. The Pexip Infinity platform cannot (and will not attempt to) create or delete AWS instances.

Configuring the bursting threshold

When enabling your platform for cloud bursting the most important decision you must make is the level at which to set the bursting threshold:

- The bursting threshold controls when your dynamic overflow nodes in your cloud service are automatically started up so that they can provide additional conferencing capacity. When the number of additional HD calls that can still be hosted in a location reaches or drops below the threshold, it triggers Pexip Infinity into starting up an overflow node in the overflow location.

  For example, setting the threshold to 5 means that when there are 5 or fewer HD connections still available in a location, an overflow node will be started up.

- When an overflow location reaches the bursting threshold i.e. the number of additional HD calls that can still be hosted on the Conferencing Nodes in the overflow location reaches the threshold, another overflow node in that location is started up, and so on.

  Note that the current number of free HD connections in the original location is ignored when deciding if the overflow location needs to overflow further — however, new calls will automatically use any available media resource that has become available within the original principal location.
- The bursting threshold is a global setting — it applies to every system location in your deployment.
- Note that it takes approximately 5 minutes for a dynamic node instance to start up and become available for conference hosting. If your principal deployment reaches full capacity, and the overflow nodes have not completed initiating, any incoming calls during this period will be rejected with "capacity exceeded" messages. You have to balance the need for having standby capacity started up in time to meet the expected demand, against starting up nodes too early and incurring extra unnecessary costs.

Manually starting an overflow node

If you know that your system will need additional capacity at a specific time due to a predictable or scheduled spike in demand, but do not want to wait for the bursting threshold to be triggered before starting up the overflow nodes, you can manually start up any of your overflow nodes.

Do not manually start an overflow node too early. If you manually start up a node more than the **Minimum lifetime** minutes before the node is needed, it will most probably get automatically stopped again before it is used.

You can start overflow nodes via the management API or via the Administrator interface:

- **Via the management API**: the cloud_node status resource can be used to list all of the available overflow nodes, the cloud_monitored_location and cloud_overflow_location resources retrieve the current load on the primary locations and any currently active overflow locations respectively, and the start_cloudnode resource can be used to manually start up any overflow node. This means that a third-party scheduling system, for example, could be configured to start up the overflow nodes via the management API approximately 10 minutes before a large conference is due to start.

  For example, let's assume that you have:
  - a regular spike in conferencing capacity demand at 9:00am every morning
  - an even usage of about 20% of that spike level during the rest of the day
  - a 30:70 ratio between your "always on" capacity and your overflow cloud capacity

  we would recommend:
  - configuring a low bursting threshold, such as 10-20% of your "always on" capacity (i.e. if your "always on" capacity is 80 HD calls, then set the bursting threshold to 12)
  - getting your scheduling system to call the API to manually start up all of your overflow cloud nodes at 8:50am on weekdays.
- **Via the Pexip Infinity Administrator interface**: go to **Status > Cloud Bursting** and select **Start** for the required nodes (the **Start** option is in the final column of the **Cloud overflow nodes** table).

Converting between overflow and "always on" AWS Conferencing Nodes

If you need to convert an existing "always on" AWS Conferencing Node into an overflow node:

1. In AWS:
   a. Apply to the instance a tag with a **Key** of **pexip-cloud** and an associated **Value** set to the **Tag value** that is shown in the **Cloud bursting** section of the **Platform > Global Settings** page.
   b. Manually stop the node instance on AWS.
2. In Pexip Infinity:
   a. Change the system location of the Conferencing Node to the overflow system location (e.g. "AWS burst").
   b. Disable the node's **Enable distributed database** setting.

      You should avoid frequent toggling of this setting. When changing this setting on multiple Conferencing Nodes, update one node at a time, waiting a few minutes before updating the next node.

If you need to convert an existing AWS overflow Conferencing Node into an "always on" node:

1. In AWS:
   a. Remove the tag with a **Key** of **pexip-cloud** from the AWS instance.
   b. Manually start the node instance on AWS.
2. In Pexip Infinity:
   a. Change the system location of the Conferencing Node to a location other than the overflow system location.
   b. Enable the node's **Enable distributed database** setting.
      You should avoid frequent toggling of this setting. When changing this setting on multiple Conferencing Nodes, update one node at a time, waiting a few minutes before updating the next node.

Managing AWS instances

This section describes the common maintenance tasks for stopping, restarting and permanently removing Conferencing Node AWS instances, and how to add ENA support to an instance.

Temporarily removing (stopping) a Conferencing Node instance

At any time you can temporarily remove a Conferencing Node instance from your Pexip Infinity platform if, for example, you do not need all of your current conferencing capacity.

To temporarily remove a Conferencing Node instance:

1. Put the Conferencing Node into maintenance mode via the Pexip Infinity Administrator interface on the Management Node:
   a. Go to **Platform > Conferencing Nodes**.
   b. Select the Conferencing Node(s).
   c. From the **Action** menu at the top left of the screen, select **Enable maintenance mode** and then select **Go**.
   While maintenance mode is enabled, this Conferencing Node will not accept any new conference instances.
   d. Wait until any existing conferences on that Conferencing Node have finished. To check, go to **Status > Live View**.
2. Stop the Conferencing Node instance on AWS:
   a. From the AWS management console, select **Instances** to see the status of all of your instances.
   b. Select the instance you want to shut down.
   c. From the **Actions** drop-down, select **Instance State > Stop** to shut down the instance.

Reinstating (restarting) a stopped Conferencing Node instance

You can reinstate a Conferencing Node instance that has already been installed but has been temporarily shut down.

To restart a Conferencing Node instance:

1. Restart the Conferencing Node instance on AWS:
   a. From the AWS management console, select **Instances** to see the status of all of your instances.
   b. Select the instance you want to restart.
   c. From the **Actions** drop-down, select **Instance State > Start** to start the instance.
2. Take the Conferencing Node out of maintenance mode via the Pexip Infinity Administrator interface on the Management Node:
   a. Go to **Platform > Conferencing Nodes**.
   b. Select the Conferencing Node.
   c. Clear the **Enable maintenance mode** check box and select **Save**.
3. Update the Conferencing Node's static NAT address, if appropriate.
   If your Conferencing Node instance was configured with an auto-assigned public IP address, it will be assigned a new public IP address when the instance is restarted.
   a. Go to **Platform > Conferencing Nodes** and select the Conferencing Node.
   b. Configure the **Static NAT address** as the instance's new public IP address.

After reinstating a Conferencing Node, it takes approximately 5 minutes for the node to reboot and be available for conference hosting, and for its last contacted status to be updated on the Management Node.

Permanently removing a Conferencing Node instance

If you no longer need a Conferencing Node instance, you can permanently delete it from your Pexip Infinity platform.

To remove a Conferencing Node instance:

1. If you have not already done so, put the Conferencing Node into maintenance mode via the Pexip Infinity Administrator interface on the Management Node:
   a. Go to **Platform > Conferencing Nodes**.
   b. Select the Conferencing Node(s).
   c. From the **Action** menu at the top left of the screen, select **Enable maintenance mode** and then select **Go**.
      While maintenance mode is enabled, this Conferencing Node will not accept any new conference instances.
   d. Wait until any existing conferences on that Conferencing Node have finished. To check, go to **Status > Live View**.
2. Delete the Conferencing Node from the Management Node:
   a. Go to **Platform > Conferencing Nodes** and select the Conferencing Node.
   b. Select the check box next to the node you want to delete, and then from the **Action** drop-down menu, select **Delete selected Conferencing Nodes** and then select **Go**.
3. Terminate the Conferencing Node instance on AWS:
   a. From the Amazon VPC console, select **Instances** to see the status of all of your instances.
   b. Select the instance you want to permanently remove.
   c. From the **Actions** drop-down, select **Instance State > Terminate** to remove the instance.

Adding ENA support to an instance

As of version 21, Pexip Infinity supports AWS instance types that use the Elastic Network Adapter (ENA). This includes C5 instance types.

If you are switching to a C5 instance and your node was originally deployed with version 20 or earlier, then you need to add ENA support to the instance, otherwise your instance will not start:

1. Ensure that your Pexip Infinity platform is running version 21 or later.
2. Shut down the instance.
3. Run the AWS CLI command:
   ```
   aws ec2 modify-instance-attribute --instance-id <instance_id> --ena-support
   ```
   See this article for the PowerShell variant.
4. Run the following command to confirm ENA support has been added:
   ```
   aws ec2 describe-instances --instance-ids <instance_id> --query "Reservations[].Instances[].EnaSupport"
   ```
   It should output:
   ```
   [ true ]
   ```
5. Power on the instance type (or change the instance type and power on).

Viewing cloud bursting status

You can view the current status of your overflow nodes and locations, and view a history of all events that have been applied to overflow nodes.

Viewing current status

Go to **Status > Cloud Bursting** to see an overview of the media load of your principal locations (that contain your "always-on" Conferencing Nodes), and whether your overflow nodes and locations are in use.

- Any issues relating to your cloud bursting deployment will also be shown on this page.
- The list of principal locations only includes those locations that are configured with a **Primary overflow location** that contains bursting nodes.
- An **approaching threshold** message is displayed in the **Available HD** connections column for the principal locations when the number of available HD connections is less than or equal to the bursting threshold plus two.

  This message changes to **bursting threshold reached** when the number of available HD connections is less than or equal to the bursting threshold (and therefore overflow nodes are started up).
- You can manually start any overflow nodes by selecting **Start** for the required node (the **Start** option is in the final column of the **Cloud overflow nodes** table).
- The status page dynamically updates every 15 seconds.

Viewing historic events

Go to **History & Logs > Conferencing Node History** to see all of the events (stop, start or running) that have been applied to overflow Conferencing Nodes and, where appropriate, the reason why the event was applied (for example if a node was shut down as there was no longer a need for the extra capacity).

© 2023 Pexip AS