



# Pexip Infinity and Amazon Web Services Deployment Guide

## Contents

<b>Introduction</b> .....	<b>1</b>
<b>Deployment guidelines</b> .....	<b>2</b>
<b>Configuring AWS security groups</b> .....	<b>4</b>
<b>Deploying a Management Node in AWS</b> .....	<b>6</b>
<b>Deploying a Conferencing Node in AWS</b> .....	<b>9</b>
<b>Dynamic bursting to the AWS cloud</b> .....	<b>12</b>
<b>Managing AWS instances</b> .....	<b>20</b>
<b>Viewing cloud bursting status</b> .....	<b>21</b>

## 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.

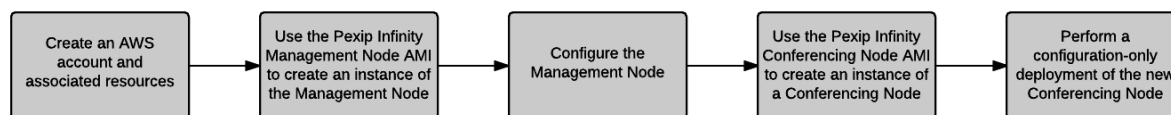
AWS enables you to scale up or down to handle changes in requirements or spikes in conferencing requirements. You can also use the AWS APIs and the Pexip Infinity management API to monitor usage and bring up / tear down Conferencing Nodes as required to meet conferencing demand.

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.

# 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.

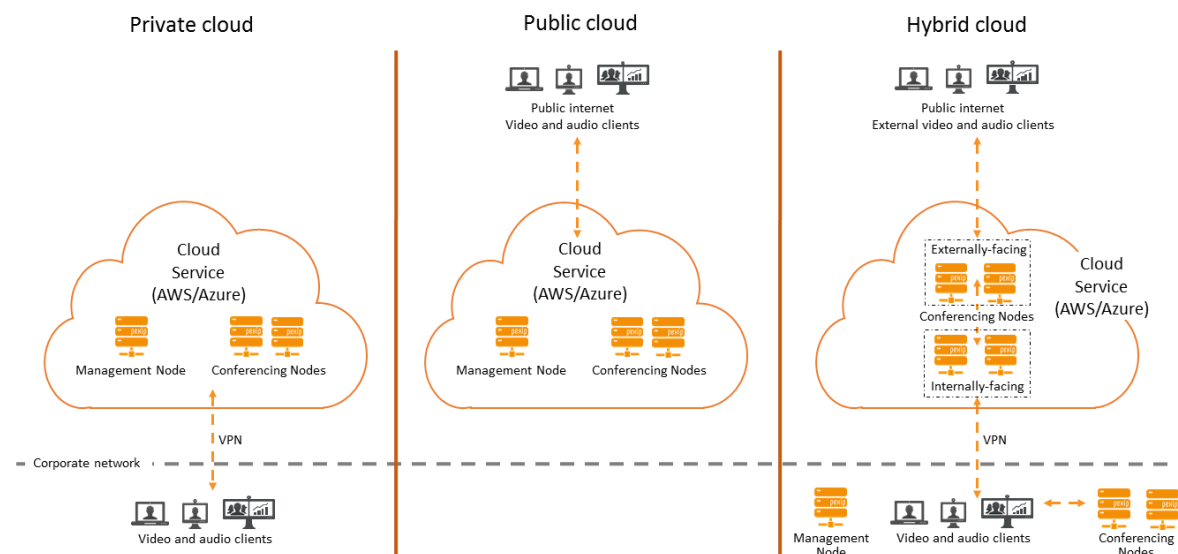
This flowchart provides an overview of the basic steps involved in deploying the Pexip Infinity platform on AWS:



## 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 from the corporate network to 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 then you must ensure that your local network allows this.
- **Hybrid cloud:** the Management Node, and optionally some Conferencing Nodes, are deployed in the corporate network. A VPN tunnel is created from the corporate network to 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).



## Limitations

The following limitations currently apply:

- All of the Pexip Infinity node instances that are hosted on AWS must be deployed in a single AWS region, so that inter-node communication between the Management Node and all of its associated Conferencing Nodes can succeed. (In a hybrid cloud deployment, some nodes may be deployed in the corporate network, but those deployed in the VPC must all be in the same AWS region.)

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.

## 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.

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

- an **m4.large** instance for a Management Node
- a **c4.2xlarge** instance for Conferencing Nodes

This should provide capacity for approximately 15 HD / 34 SD / 175 audio-only calls per Conferencing Node.

Note that m3.large and c3.2xlarge can be used instead if m4.large and c4.2xlarge are not available in your region. Larger instance types (such as c4.4xlarge and c4.8xlarge) may also be used for a Conferencing Node, but the call capacity does not increase linearly so these may not represent the best value.

## IP addressing

Within a VPC, private IP addresses may be allocated dynamically (using DHCP) or statically, by defining an instance's IP address at launch time. After a private IP address has been assigned to an instance, it will remain 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. To associate the instance's public IP address with the node, configure that public IP address as the node's **Static NAT address** (via **Platform Configuration > 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 (one Management Node and one or more associated Conferencing Nodes).

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>.

To look at the steps taken in setting up an example lab deployment of a Management Node in AWS, see <http://www.graham-walsh.com/2016/01/deploying-pexip-management-node-in-amazon-web-services/>, and to see an example of deploying a Conferencing Node in AWS, see <http://www.graham-walsh.com/2016/01/deploying-pexip-conference-node-in-amazon-web-services/>.

## 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 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](#).

# 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 **Launch Instance**.  
This launches the wizard in which you will select and configure your image.
3. Complete Step 1: Choose an Amazon Machine Image (AMI):
  - a. Select **Community AMIs**.
  - b. Search the Community AMIs section for "Pexip".
  - c. Select **Pexip Infinity Management Node <version> build <build\_number>** where **<version>** is the software version you want to install.
4. Complete Step 2: Choose an Instance Type:
  - a. For deployments of up to 30 Conferencing Nodes, we recommend using an **m4.large** 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 <i>Auto-assign</i> 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.)

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

6. Complete Step 4: Add Storage:
  - a. Accept the default settings (the Pexip AMI will have set these defaults appropriately for a Management Node).
  - b. Select **Next: Tag Instance**.
7. Complete Step 5: Tag Instance:
  - 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.

AWS
Services
Edit

Ireland
Support

1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Tag Instance
6. Configure Security Group
7. Review

### Step 7: Review Instance Launch

AMI Details

**Pexip Infinity Management Node 11.0.0 (build 26902.0.0) - ami-e5d67696**  
Pexip Infinity Management Node 11.0.0 (build 26902.0.0)  
Root Device Type: ebs    Virtualization type: hvm

Edit AMI

Instance Type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
m4.large	6.5	2	8	EBS only	Yes	Moderate

Edit instance type

Security Groups

Security Group ID	Name	Description
sg-0e09a56a	pexip_group	Pexip security group

Edit security groups

All selected security groups inbound rules

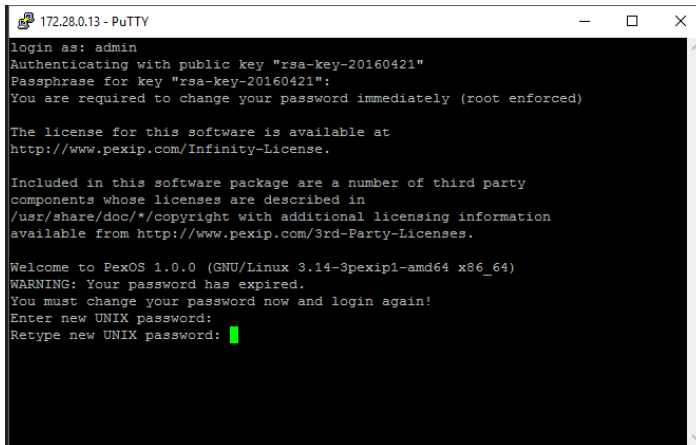
Security Group ID	Type	Protocol	Port Range	Source
sg-0e09a56a	Custom TCP Rule	TCP	1720	0.0.0.0/0
sg-0e09a56a	SSH	TCP	22	10.0.0.0/8
sg-0e09a56a	Custom TCP Rule	TCP	8443	10.0.0.0/8
sg-0e09a56a	Custom TCP Rule	TCP	5061	0.0.0.0/0
sg-0e09a56a	Custom UDP Rule	UDP	1719	0.0.0.0/0
sg-0e09a56a	Custom Protocol	ESP (50)	All	sg-0e09a56a (pexip_group)
sg-0e09a56a	Custom TCP Rule	TCP	33000 - 49999	0.0.0.0/0
sg-0e09a56a	Custom UDP Rule	UDP	500	sg-0e09a56a (pexip_group)
sg-0e09a56a	HTTPS	TCP	443	0.0.0.0/0
sg-0e09a56a	All ICMP	All	N/A	10.0.0.0/8
sg-0e09a56a	Custom UDP Rule	UDP	40000 - 49999	0.0.0.0/0
sg-0e09a56a	Custom TCP Rule	TCP	5060	0.0.0.0/0

Cancel
Previous
Launch

- 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. You must now 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.



```

172.28.0.13 - PuTTY
login as: admin
Authenticating with public key "rsa-key-20160421"
Passphrase for key "rsa-key-20160421":
You are required to change your password immediately (root enforced)

The license for this software is available at
http://www.pexip.com/Infinity-License.

Included in this software package are a number of third party
components whose licenses are described in
/usr/share/doc/*/copyright with additional licensing information
available from http://www.pexip.com/3rd-Party-Licenses.

Welcome to PexOS 1.0.0 (GNU/Linux 3.14-3pexipl-amd64 x86_64)
WARNING: Your password has expired.
You must change your password now and login again!
Enter new UNIX password:
Retype new UNIX password:
  
```

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:  
Running Pexip installation wizard...  
[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:
      - i. Accept the defaults for the **IP address**, **Network mask** and **Gateway** settings.
      - ii. Enter your required **Hostname** and **Domain suffix** for the Management Node.
      - iii. Configure one or more **DNS servers** and **NTP servers**. You must override the default values if it is a private deployment.
      - iv. Set the **Web administration username** and **password**.
      - v. Select whether to **Enable incident reporting** and whether to **Send deployment and usage statistics to Pexip**.  
 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 now configure your Pexip Infinity platform licenses, VMRs, aliases, locations etc, and add Conferencing Nodes.

## Deploying a Conferencing Node in AWS

After deploying the Management Node 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. 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.
2. Search the Community AMIs section for the relevant Pexip Infinity Conferencing Node AMI.
3. 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.
4. After the instance has booted, perform a configuration-only deployment on the Management Node to inform it of the new Conferencing Node.
5. Upload the resulting XML document to the new Conferencing Node.
6. Configure the Conferencing Node's static NAT address, if you have assigned a public IP address to the instance.

These steps are described below in more detail.

### Task breakdown

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 **Launch Instance**.  
This launches the wizard in which you will select and configure your image.
3. Complete Step 1: Choose an Amazon Machine Image (AMI):
  - a. Select **Community AMIs**.
  - b. Search the Community AMIs section for "Pexip".
  - c. Select **Pexip Infinity Configuration Node <version> build <build\_number>** where **<version>** is the software version you want to install.
4. Complete Step 2: Choose an Instance Type:
  - a. We recommend using a **c4.2xlarge** instance type for the Conferencing 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	<p>Enable or disable this option according to whether you want the node to be reachable from a public IP address. You must assign a public IP address to the Conferencing Node if you want that node to be able to host conferences and be accessible from devices in the public internet.</p> <p>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.</p>
Primary IP	<p>Either leave as <i>Auto-assign</i> or, if required, specify your desired IP address.</p> <p>(AWS reserves the first four IP addresses and the last one IP address of every subnet for IP networking purposes.)</p>
Tenancy	Select <i>Dedicated - Run a Dedicated instance</i> .

- b. Select **Next: Add Storage**.
6. Complete Step 4: Add Storage:
  - a. Accept the default settings (the Pexip AMI will have set these defaults appropriately for a Conferencing Node).
  - b. Select **Next: Tag Instance**.
7. Complete Step 5: Tag Instance:
  - 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.

AWS

Services

Edit

Ireland

Support

1. Choose AMI

2. Choose Instance Type

3. Configure Instance

4. Add Storage

5. Tag Instance

6. Configure Security Group

7. Review

### Step 7: Review Instance Launch

AMI Details

Pexip Infinity Conferencing Node 11.0.0 (build 26902.0.0) - ami-60dd7d13

Pexip Infinity Conferencing Node 11.0.0 (build 26902.0.0)

Root Device Type: ebs    Virtualization type: hvm

Edit AMI

Instance Type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
c4.2xlarge	31	8	15	EBS only	Yes	High

Edit instance type

Security Groups

Security Group ID	Name	Description
sg-0e09a56a	pexip_group	Pexip security group

Edit security groups

All selected security groups inbound rules

Security Group ID	Type	Protocol	Port Range	Source
sg-0e09a56a	Custom TCP Rule	TCP	1720	0.0.0.0/0
sg-0e09a56a	SSH	TCP	22	10.0.0.0/8
sg-0e09a56a	Custom TCP Rule	TCP	8443	10.0.0.0/8
sg-0e09a56a	Custom TCP Rule	TCP	5061	0.0.0.0/0
sg-0e09a56a	Custom UDP Rule	UDP	1719	0.0.0.0/0
sg-0e09a56a	Custom Protocol	ESP (50)	All	sg-0e09a56a (pexip_group)
sg-0e09a56a	Custom TCP Rule	TCP	33000 - 49999	0.0.0.0/0
sg-0e09a56a	Custom UDP Rule	UDP	500	sg-0e09a56a (pexip_group)
sg-0e09a56a	HTTPS	TCP	443	0.0.0.0/0
sg-0e09a56a	All ICMP	All	N/A	10.0.0.0/8
sg-0e09a56a	Custom UDP Rule	UDP	40000 - 49999	0.0.0.0/0
sg-0e09a56a	Custom TCP Rule	TCP	5060	0.0.0.0/0

Cancel

Previous

Launch

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:
  - a. Log in to the Pexip Infinity Administrator interface on the Management Node.
  - b. Go to Platform Configuration > Conferencing Nodes.
  - c. Select Add Conferencing Node.
  - d. For deployment type, choose *Generic (configuration-only)*.

- e. Enter the details of the new Conferencing Node, including:

IPv4 address	Enter the Private IP address that AWS has assigned to the new Conferencing Node.
Network mask	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 IP address	The gateway address 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).

- f. Select **Finish**.

- g. Select **Download Conferencing Node Configuration** and save the XML configuration file.

A zip file with the name **pexip-<hostname>.<domain>.xml** will be downloaded.

14. You must now upload the XML configuration file to the new Conferencing Node:

- a. Browse to **https://<conferencing-node-private-ip>:8443/** and use the form provided to upload the XML configuration file to the Conferencing Node VM.

i. Select **Choose File** and select the XML configuration file.

ii. Select **Upload**.

- b. The Conferencing Node will apply the configuration and then reboot. When it has rebooted, it will connect to the Management Node.

You can close the browser window used to upload the file.

15. If you want the node to have a persistent public IP address you can assign an Elastic IP address to the Conferencing Node. To do this, use the **Elastic IPs** option in the Amazon VPC console.

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.

16. Configure the Conferencing Node's static NAT address, if you have assigned a public IP address to the instance:

- a. Log in to the Pexip Infinity Administrator interface on the Management Node.

- b. Go to **Platform Configuration > Conferencing Nodes** and select the Conferencing Node.

- c. Configure the **Static NAT address** as the instance's public IP address (either the auto-assigned public address or the Elastic IP address as appropriate).

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 is available, the Management Node will report the status of the Conferencing Node as having a last contacted and last updated date of "Never".)

## 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.

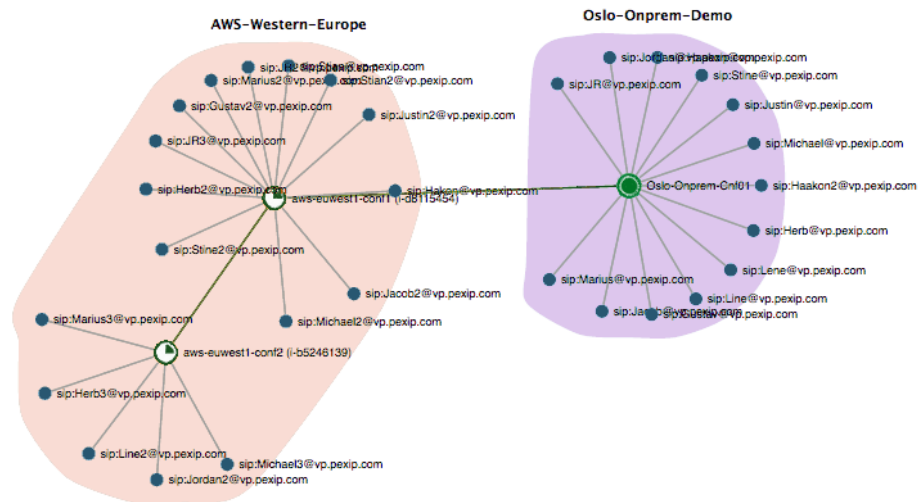
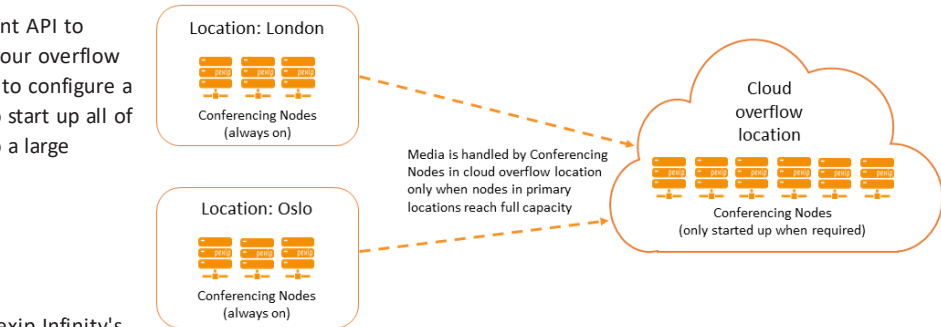
You can also use the management API to monitor and manually start up your overflow nodes, if for example, you want to configure a third-party scheduling system to start up all of your additional capacity prior to a large conferencing event starting.

## How it works

Dynamic bursting builds upon Pexip Infinity's standard location capacity overflow logic, which is used to define which overflow location's nodes are used when a particular location reaches its capacity. The dynamic bursting functionality adds to this by automatically starting up those additional AWS cloud nodes when a location is **approaching** full capacity, so that those overflow nodes will be available if required.

After you have deployed your overflow Conferencing Nodes in AWS, and enabled cloud bursting and configured your bursting threshold in Pexip Infinity, everything is then controlled automatically by the Pexip Infinity Management Node:

- Whenever the available capacity in a system location containing your primary (always on) Conferencing Node hits or drops below your configured threshold, a Conferencing Node in a system location containing overflow Conferencing Nodes is automatically started up. That new node can then start handling any additional conferencing requirements if the original location reaches full capacity.



- When an overflow Conferencing Node starts to fill up and reaches its own bursting threshold, a further overflow Conferencing Node is started, and so on. The conference graph (right) shows a conference hosted in an on-premises Oslo location that has burst onto two Conferencing Nodes in the "AWS Western Europe" overflow location.
- When call levels subside and an overflow AWS Conferencing Node is no longer hosting conferences and is no longer required, the node is automatically shut down again.

This sequence of events is explained in more detail in a [worked example](#).

## Configuring your system for dynamic bursting

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.

## 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 those nodes in Pexip Infinity as the overflow destination for your primary (always on) Conferencing Nodes:

1. In Pexip Infinity, configure a new "overflow" system location 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 and 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 AWS instance to be used for conference bursting a tag with a Key of **pexip-cloud** and an associated Value set to the AWS tag value that is shown at the bottom of the **Platform Configuration > Global Settings** page.  
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 Configuration > Global Settings**, and enable cloud bursting and configure your bursting threshold:

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 AWS when in need of extra capacity.
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 <b>Identity And Access Management</b> in step 2 above.
Bursting threshold	The bursting threshold controls when your overflow nodes in AWS 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 <a href="#">Configuring the bursting threshold</a> for more information.
AWS tag name and AWS tag value	These read-only fields indicate the tag name (always <b>pexip-cloud</b> ) and associated tag value (the hostname of your Management Node) that you must assign to each of your AWS instances that are to be used for dynamic bursting.

5. Go to **Platform Configuration > Locations** and configure the system locations that contain your primary (always on) Conferencing Nodes so that they will overflow to your new "AWS burst" location.  
When configuring these locations, you must set the **Primary overflow location** to the bursting location containing your overflow nodes. (Automatic bursting, and the stopping and starting of overflow nodes only applies to the **Primary overflow location**; the **Secondary overflow location** can only be used for standard overflow i.e. to other "always on" nodes.)

## 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 dashboard, select **Identity and Access Management**.
2. Select **Users** and create a new user on behalf of the Pexip platform e.g. username "pexip". Ensure that "Generate an access key for each user" is selected.
3. Either download the user credentials or select **Show User Security Credentials** 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.)
4. Select **Policies** and then **Create Your Own Policy** to set up the access policy for the overflow nodes:
  - a. Enter a **Policy Name** and **Description**.
  - b. For the **Policy Document**, 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": "<management-node-hostname>"
        }
      }
    }
  ]
}
```

- c. The only element of this policy document that you need to change is to replace **<management-node-hostname>** with the hostname of your Management Node (as shown via **Platform Configuration > Management Node**). This is the same name as you set as the value of the **pexip-cloud** tag that you assigned to your AWS overflow Conferencing Nodes.
  - d. Select **Validate Policy** and then (if valid) **Create Policy**.
5. Attach your "pexip" user to your policy:
    - a. From the **Policies** page, select the checkbox next to your policy (you can filter on Customer Managed Policies if necessary).
    - b. From the **Policy Actions** dropdown select **Attach**, select the checkbox next to your **pexip** user and select **Attach Policy**.
- 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 overflow nodes in AWS 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 are not taken into account when seeing if the overflow location needs to overflow further — however, new calls will automatically use any available media resource within those original primary locations that has become available.
- 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 Conferencing Node instance in AWS to start up and become available for conference hosting. If your primary 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:

- **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).

## Viewing cloud bursting status

Go to **Status > Cloud Bursting** to see an overview of the media load of your primary 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 primary locations only includes those system 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 primary 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.

## Deployment guidelines and additional information

- An overflow Conferencing Node is automatically stopped when it becomes idle (no longer hosting any conferences). However, a node is never stopped until it has been running for at least 50 minutes. This is because AWS charges by the hour, so when a node is started up it is more efficient to leave it running for 50 minutes — even if it is never used — as that capacity can remain on immediate standby for no extra cost.
- A location can trigger an overflow node to start up only if there is some existing media load on that location. This is to ensure that incidents such as a temporary network interruption do not inadvertently trigger bursting.
- We recommend that you do not mix your primary (always on) Conferencing Nodes and your bursting nodes in the same system location.
- As per standard AWS configuration guidelines, all of your AWS Conferencing Nodes must be in the same AWS region.
- Typically you should assign all of your overflow Conferencing Nodes to a single overflow system location so that all of your primary locations can make use of the same pool of cloud overflow nodes.

However, if you expect a single conference to require more than three overflow nodes, then you could configure a second cloud overflow location (containing a different set of AWS overflow nodes) and then configure half of your primary locations to overflow to the first cloud overflow location, and the other half of your primary locations to overflow to the second cloud overflow location. In this case, both overflow locations will act (start up and shut down nodes as appropriate) independently of each other.

- Do not configure your call control system / DNS to route calls directly to your overflow nodes.
- No specific licenses are required for your overflow nodes, but you must ensure that your overall system has sufficient licenses to meet peak conference capacity demand.
- If you need to convert an existing "always on" AWS Conferencing Node into an overflow node:
  - In AWS:
    - i. Apply a tag with a **Key** of **pexip-cloud** and an associated **Value** set to the **AWS tag value** that is shown at the bottom of the **Platform Configuration > Global Settings** page.
    - ii. Manually stop the node instance on AWS.
  - In Pexip Infinity:
    - i. Change the system location of the Conferencing Node to the overflow system location (e.g. "AWS burst").
    - ii. Disable the node's **Enable distributed database** setting. After a node has been deployed, this setting can only be changed via the [Management configuration API](#) using the `worker_vm` resource.
- If you need to convert an existing AWS overflow Conferencing Node into an "always on" node:
  - In AWS:
    - i. Remove the tag with a **Key** of **pexip-cloud** from the AWS instance.
    - ii. Manually start the node instance on AWS.
  - In Pexip Infinity:
    - i. Change the system location of the Conferencing Node to a location other than the overflow system location.
    - ii. Enable the node's **Enable distributed database** setting. After a node has been deployed, this setting can only be changed via the [Management configuration API](#) using the `worker_vm` resource.

## Troubleshooting

### Checking the administrator and support logs

All log messages that are explicitly related to cloud bursting, such as starting up or shutting down overflow nodes, are tagged with a log module Name of `administrator.apps.cloudbursting`.

All other log messages related to those overflow nodes or the conferences they are hosting are reported in the same manner as per standard behavior.

### No AWS nodes appear in the Cloud overflow nodes area of the Status > Cloud Bursting page

Check that the AWS instances have been assigned the `pexip-cloud` tag and that the tag value is set to the Management Node hostname.

### Instance found but no corresponding Conferencing Node

You may see a status issue "Instance <name> (with IP <address>) was found, but no corresponding Conferencing Node has been configured".

This occurs when Pexip Infinity detects an AWS instance with a tag matching your system's hostname but there is no corresponding Conferencing Node configured within Pexip Infinity.

This message can occur temporarily in a normal scenario when deploying a new Conferencing Node and you have set up the VM instance in AWS but you have not yet deployed the Conferencing Node in Pexip Infinity. In this case, the issue will disappear as soon as you have deployed the Conferencing Node.

### Connectivity errors in the administrator log

You may see connectivity errors in the administrator log while overflow nodes are being started/stopped. This is normal behavior.

### Not authorized to perform an operation

A "Not authorized to perform an operation on <instance ID or region name>. Check the policy created for the AWS user." means that there is a problem with the AWS policy document, or the AWS user is not attached to the policy. See [Configuring an AWS user and policy for controlling overflow nodes](#) for more information.

## Detailed example of the overflow process

This sequence of actions listed below shows how the process of starting and stopping overflow nodes is managed. It assumes an example scenario where:

- the system is configured with a bursting threshold of 5
- there is a single primary location (London) containing 2 "always on" Conferencing Nodes with a total location capacity of 40 connections
- the primary location is configured to overflow to the cloud overflow location "AWS burst"
- the "AWS burst" location contains 2 overflow nodes, each with a capacity of 20 connections.

Primary location: London			Cloud overflow location: AWS burst		
Action	Call capacity remaining	Dynamic bursting activity	Overflow node A status	Overflow node B status	Call capacity remaining
1. No conferences in progress	40	<i>none</i>	Not running	Not running	
1. A conference starts and has 33 participants	6*	<i>none</i>			
2. 1 more participant joins	5	Primary location threshold reached	Node A starting		
3.	5	Overflow capacity becomes available	Running		20

Primary location: London			Cloud overflow location: AWS burst		
Action	Call capacity remaining	Dynamic bursting activity	Overflow node A status	Overflow node B status	Call capacity remaining
4. 5 more participants join	0	<i>none</i>	↓		20
5. 1 more participant joins	0	Call media handled by overflow node A	↓		18*
6. 12 more participants join	0	All new participants handled by overflow node A	↓		6
7. 1 more participant joins	0	Overflow location threshold reached	↓	Node B starting	5
8.	0	Additional overflow capacity becomes available	↓	Running	25
9. 7 more participants join	0	All new participants handled by overflow nodes A and B	↓	↓	17*
10. Conference ends	40	Overflow nodes remain available for further bursting if required	↓	↓	40
11. A new conference starts with 25 participants	14*	<i>none</i>	↓	↓	40
12.		Overflow node A is unused and has been running for 50 minutes	Shutting down	↓	20
13.		Overflow node B is unused and has been running for 50 minutes	Not running	Shutting down	0
14. Conference ends	40		Not running	Not running	

\* a Conferencing Node reserves 1 HD connection for the backplane to other nodes in the same conference

Note that if other conferences were running on any of the nodes in these locations, they would consume call capacity and bursting would be triggered in exactly the same way when the remaining capacity within the location reached the threshold.

## Managing AWS instances

This section describes the common maintenance tasks for [stopping](#), [restarting](#) and [permanently removing](#) Conferencing Node AWS instances.

### 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 Configuration > Conferencing Nodes**.
  - b. Select the Conferencing Node.
  - c. Select the **Enable maintenance mode** check box and select **Save**.  
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 Configuration > 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 Configuration > 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 Configuration > Conferencing Nodes**.
  - b. Select the Conferencing Node.
  - c. Select the **Enable maintenance mode** check box and select **Save**.  
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 Configuration > 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.

## Viewing cloud bursting status

Go to **Status > Cloud Bursting** to see an overview of the media load of your primary 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 primary locations only includes those system 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 primary 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.