Triton Image Allowlist

Table of contents

> ⚠ **Attention**
>
> NVIDIA Triton Management Service (TMS) will reach the end of life on July 31, 2024. The version 1.4.0 is the last release.

Overview

The Triton image allowlist is a management feature used to controls which Triton images can be used to created Triton pools and bespoke Triton instances. This feature gives administrators some basic controls over which images can and cannot be used when creating Triton instances. Simply put, if an image is not in the allowlist, it cannot be used to create a new Triton instance. The Triton allowlist service is used to inspect and modify the allowlist.

Initially, the allowlist only contains the default Triton image, as configured during installation. This can be seen by running `tmsctl allowlist list`. Right after installing TMS, it should look like this:

```
$ tmsctl allowlist list
nvcr.io/nvidia/tritonserver:23.09-py3
```

In this state, new Triton instances can only be created with `nvcr.io/nvidia/tritonserver:23.09-py3` as the Triton image. For example, trying to use `nvcr.io/nvidia/tritonserver:23.08-py3` will fail.

```
$ tmsctl lease create --triton-image nvcr.io/nvidia/tritonserver:23.08-py3 -m
"name=$MODEL_NAME,uri=$MODEL_URI"
fatal: Requested Triton container image ("nvcr.io/nvidia/tritonserver:23.08-py3") is
unreachable or not provided in a supported format. Unreachable container images
either do not exist or require privileges not granted to the server.
(triton_options_bespoke.triton.container_image @ Acquire)
```

New entries can be added via `tmsctl allowlist add`.

```
$ tmsctl allowlist add nvcr.io/nvidia/tritonserver:23.08-py3
Added nvcr.io/nvidia/tritonserver:23.08-py3
$ tmsctl allowlist list
nvcr.io/nvidia/tritonserver:23.08-py3
nvcr.io/nvidia/tritonserver:23.09-py3
```

After running the above, we can create new bespoke Triton instances and Triton pools specifying `nvcr.io/nvidia/tritonserver:23.08-py3` as the image.

```
$ tmsctl lease create --triton-image nvcr.io/nvidia/tritonserver:23.08-py3 -m
"name=$MODEL_NAME,uri=$MODEL_URI"
Lease da21b2c0e68b49ffa8f0f6db0b030128
State: Valid
Expires: 2023-10-18T15:43:53Z
Triton: triton-6d8c9d13.tmsns.svc.cluster.local
<nvcr.io/nvidia/tritonserver:23.08-py3>
Models:
Name Url Status
<model_name> <model_url> Ready
```

Entries can be removed via `tmsctl allowlist rm`.

```
$ tmsctl allowlist rm nvcr.io/nvidia/tritonserver:23.10-py3
Removed nvcr.io/nvidia/tritonserver:23.10-py3
$ tmsctl allowlist list
nvcr.io/nvidia/tritonserver:23.09-py3
```

After running the above, any attempts to create new bespoke Tritons or Triton pools specifying `nvcr.io/nvidia/tritonserver:23.08-py3` as the image will fail.

Further details about the individual operations are given below.

Triton Image Allowlist Operations

Following are the list of operations with the Triton Allowlist Service:

1. `TritonAllowlist/Append` appends a Triton container image to the list containing the allowed Triton container images.

   The server will return success or failure depending on whether the requested image could be added to the allowlist.

   Attempting to add images which are already present in the allowlist will not result in any changes.

2. `TritonAllowlist/List` lists the allowed Triton container images.

   This RPC begins streaming a response once the request has been received.

   Each response message contains a Triton image that belongs to the list.

3. `TritonAllowlist/Remove` removes a Triton container image from the list containing the allowed Triton container images.

   The server will return success or failure depending on whether the image could be removed from the allowlist.