

Confidential ACI

Security Policy aka Execution Policy

Ken Gordon (at microsoft.com)

Many slides by Matthew Johnson

Practical, deployed, integrated into Azure, but also flexible.

[Confidential containers on Azure Container Instances - Azure Container Instances | Microsoft Learn](#)

[microsoft/hcsshim: Windows - Host Compute Service Shim \(github.com\)](#)

[microsoft/confidential-aci-examples: A collection of examples and tests to run on Confidential Azure Container Instances \(github.com\)](#)

LCOW

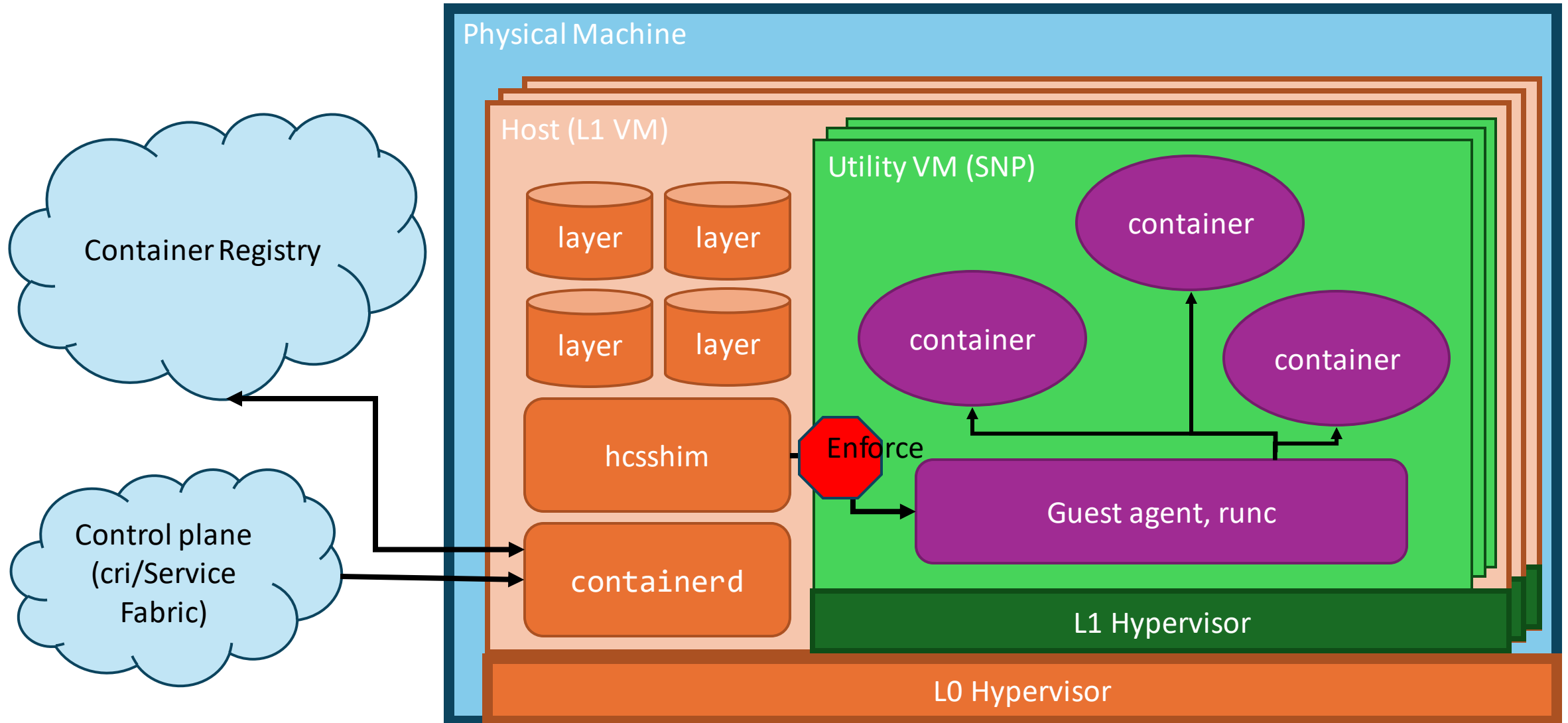
- Confidential ACI provides serverless Confidential Containers in Azure using LCOW (Linux Containers on Windows).
- Pods (aka Container Groups) owned by a customer are run in a Hyper-V isolated VM (UVM) solely for use by that customer. This VM is encrypted and protected against the host using AMD SEV-SNP.
- Containerd runs on the (Windows) host and fetches regular container images.

Delta for confidential

- The Linux guest OS (in the UVM) is minimalist and only runs the guest agent.
- An execution policy (in rego) controls what the host/control plane can ask the UVM to do.
- The guest OS is measured by the PSP and the measurement and hash of the policy is in the hardware attestation reports.
- Image layers have a Merkle tree, regenerated locally.
- “Scratch space” VHDs used for RW parts is encrypted.
- Most users will use tooling to generate the policy. Eg [az confcom | Microsoft Learn](#)

[illegible]

LCOW Infrastructure in Azure



A quick example to set the scene.

```
package policy

import future.keywords.every
import future.keywords.in

api_version := "0.10.0"
framework_version := "0.2.3"

fragments := [
  {
    "feed": "mcr.microsoft.com/aci/aci-cc-infra-fragment",
    "includes": [
      "containers"
    ],
    "issuer":
      "did:xs09:0:sha256:I__iuL25oXEVfdTP_0BLX_eT1RPHbCQ_ECBQfYZpt9s::eku:1.3.6.1.4.1.311.76.59.1.3",
    "minimum_svn": "1"
  }
]

containers := [
  {
    "allow_elevated": true,
    "allow_stdio_access": true,
    <snip>
    "command": [
      "/skr-debug.sh"
    ],
    "env_rules": [
      {
        "pattern": "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin",
        "required": false,
        "strategy": "string"
      }
    ],
    <snip>
  },
  "exec_processes": [
    <snip>
  ],
  "id": "parnawesteuroperegistry.azurecr.io/kenskr:2.0",
  "layers": [
    <snip>
    "1f8bc802e35608b7193f4c7b3632c17970d7b764889609c541566ef10d26c6f6"
  ],
  "mounts": [
    {
      "destination": "/etc/resolv.conf",
      "options": [
        "rbind",
        "rshared",
        "rw"
      ],
      "source": "sandbox:///tmp/atlas/resolvconf/."+ ,
      "type": "bind"
    }
  ],
  "no_new_privileges": false,
  "seccomp_profile_sha256": "",
  "signals": [],
  "user": {
    "group_idnames": [
      {
        "pattern": "",
        "strategy": "any"
      }
    ],
    "umask": "0022",
    "user_idname": {
      "pattern": "",
      "strategy": "any"
    }
  }
]
```

What did policy ever do for Confidential Computing?

Necessary features to ensure a workload is confidential

Protect sensitive memory - encryption/isolation

Be sure we are running in a genuine environment - attestation

Prevent interference - integrity

Control what the environment can do – policy

Release secret gated on complying with the above – relying party

How did we live without it?

Types of Policy:

Implicit policy – attested code can only do what the code can do – e.g. SGX

Explicit policy – attested code uses rules to control its behaviour

Configuration – data which is policy really, eg “only run the next blob if the blob’s hash is X”

Here we are talking about explicit policy expressed as Rego for Confidential Containers “Security Policy”

Where is the magic?

Confidential Containers run inside an encrypted UVM (Utility VM) via an agent

Before UVM start HOST_DATA (an immutable hardware register) is set to sha256(policy)

Next the UVM is measured by AMD hardware

Then it starts, is given the policy and checks the actual policy matches HOST_DATA

In the process of starting containers (and other ops) the agent **enforces the policy**

A container obtains an **attestation report** from the hardware (ignoring key wrapping)

It sends the report to a relying party

The relying party checks that the **UVM measurement** and the **policy hash** it expects are those in the **hardware report**, if so it releases the secret (ignoring mHSM details)

The policy is essential and must prevent unintended changes within the UVM:

- Wrong container
- Wrong container configuration

Attestation.

The workload must provide an attestation report and some collateral to a relying party

The attestation report is a regular AMD SEV-SNP one as illustrated earlier

AMD provides a certificate chain to validate that the attestation report is genuine

Microsoft provides a COSE Sign1 document that allows a relying party to check the hardware UVM measurement is of a genuine image.

Optionally the actual security policy can be provided. For example, CCF will record the whole policy in an immutable log.

The relying party can then check that the UVM measurement is to be trusted (via the COSE Sign1 document or an explicit value) AND that the policy hash (HOST_DATA) is as expected.

Can we explain it?

- Talk about rules or white lists
- Say “allows” about things like container lists
- Make clear the differences
 - Configuration – what ought to be done
 - Policy/rules – what is allowed to be done
- **Checking** that rules are being respected – **correct UVM**
- That the **correct rules** are in place
- The check is for **secret release** NOT running the code

A UVM cannot check itself is genuine since compromised UVM could run/do anything

Enforcement

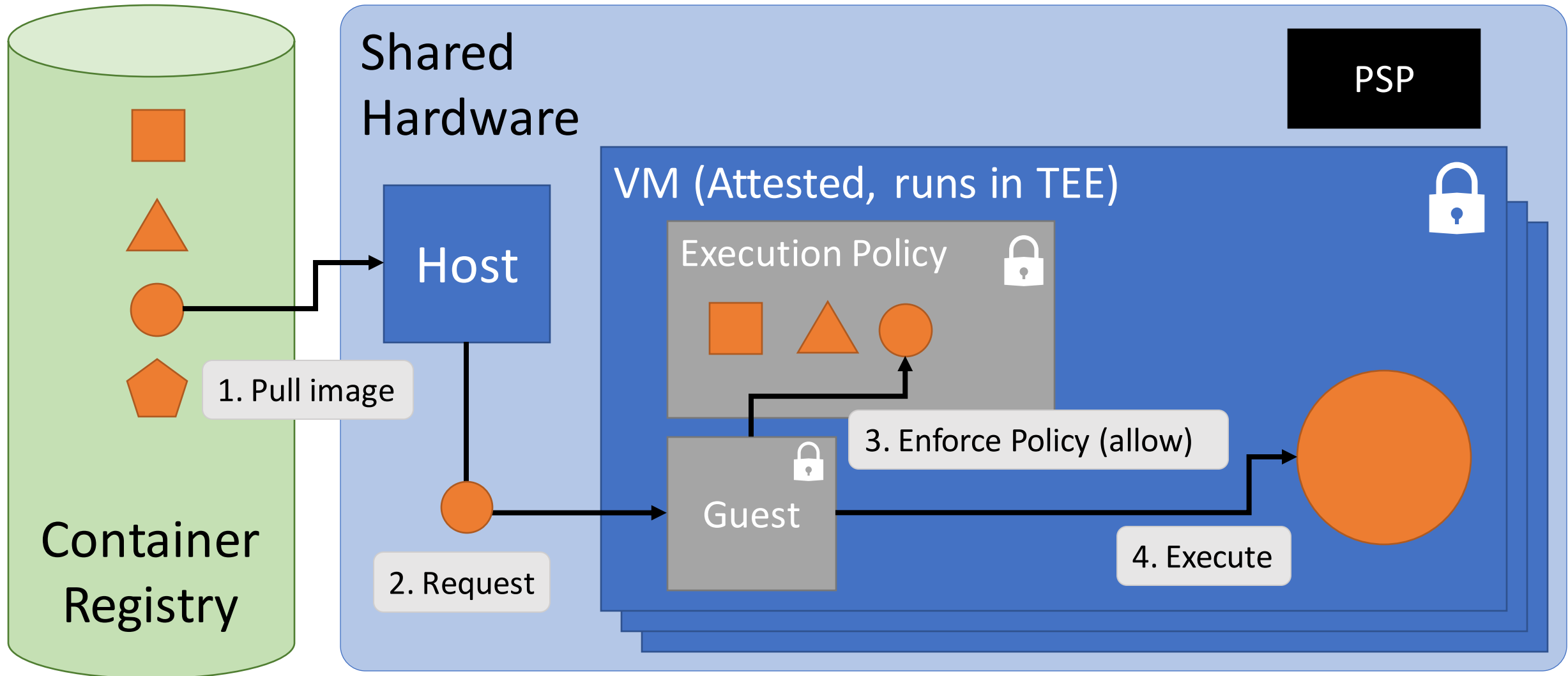
The aim of the policy is to prevent an attacker modifying the state of the UVM or Container to gain access to confidential data

For example, it constrains the containers being loaded to the trusted set specified by the user.

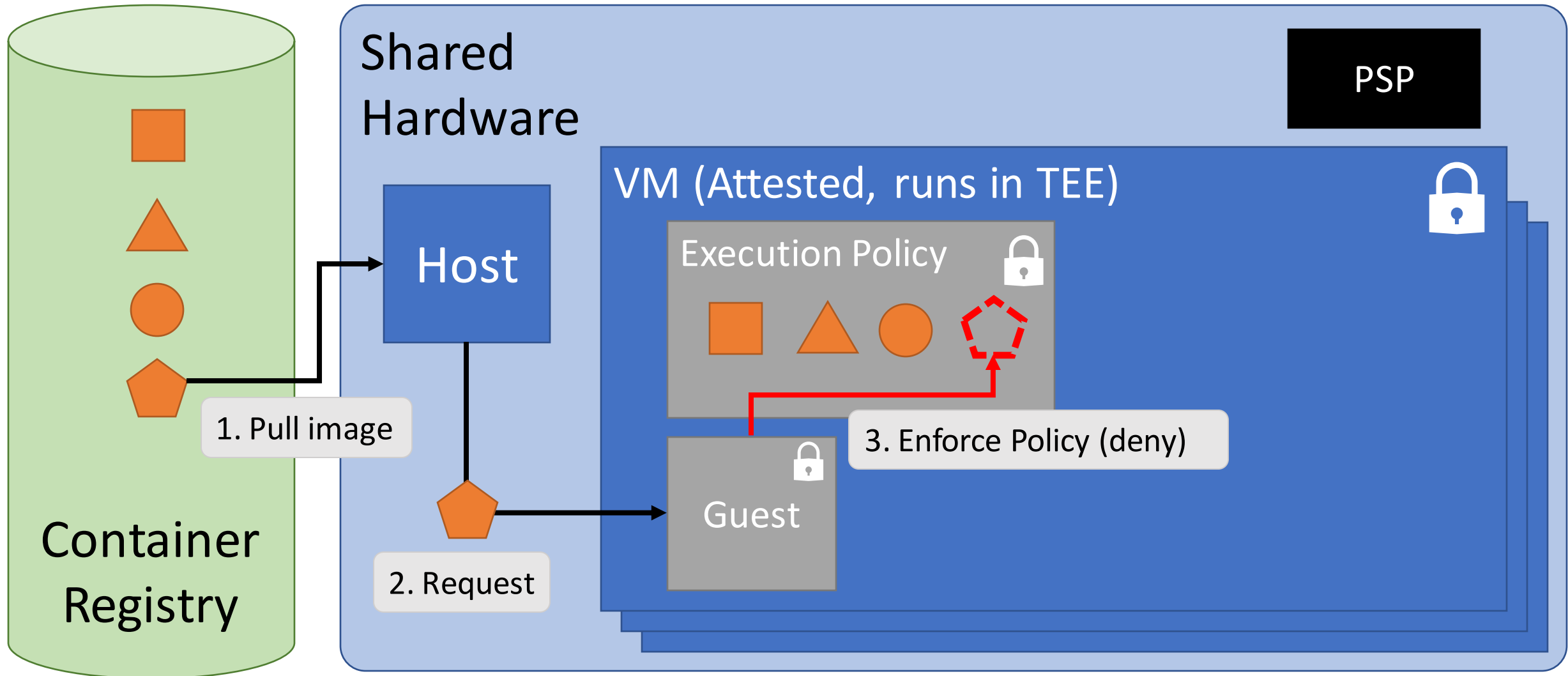
The agent code checks requests against the policy at “enforcement points”
The enforcement points must cover all significant modifications of the system state

Container runtime agents have a narrow API which enables enforcement to be straight-forward

Execution Policy: Allow



Execution Policy: Deny



What should be enforced?

Essentially the VM's external surface.

e.g.,
don't let the host mount
/evil in place of /bin,

don't allow container
evil.acurecr.io/steal_it:latest
to load

but expressed as positives

- Mounting devices
 - Mounting overlays
 - Creating containers
 - Command
 - Environment variables
 - Mounts
 - Processes
 - Logging
 - Signals
 - stdio access
-
- External (UVM) processes
 - External logging
 - Plan9 mounts
 - Getting properties
 - Dumping stacks
 - Dropping environment variables
 - Dropping capabilities
 - Security context
 - RunAsUser, RunAsGroup
 - Capabilities
 - Seccomp
 - NoNewPrivileges

```
package policy
```

```
import future.keywords.every
import future.keywords.in
```

```
api_version := "0.10.0"
framework_version := "0.2.3"
```

Confidential ACI Rego policy

```
fragments := [
{
  "feed": "mcr.microsoft.com/aci/aci-cc-infra-fragment",
  "includes": [
    "containers",
    "fragments"
  ],
  "issuer": "did:x509:0:sha256:I__iuL25oXEVFdTP_aBLx_eT1RPHbCQ_ECBQfYZpt9s::eku:1.3.6.1.4.1.311.76.59.1.3",
  "minimum_svn": "1"
}
]
```

```
containers := [
{
  "allow_elevated": false,
  "allow_stdio_access": true,
  "capabilities": { "ambient": [], "bounding": [ "CAP_AUDIT_WRITE", <snip> "CAP_SYS_CHROOT" ] },
  "command": [
    "/skr-debug.sh"
  ],
  "env_rules": [
    { "pattern": "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin", "required": false, "strategy": "string" },
    { "pattern": "TERM=xterm", "required": false, "strategy": "string" },
    <snip>
  ],
  "exec_processes": [],
  "id": "parmawesteuoperegistry.azurecr.io/kenskr:2.0",
  "layers": [
    "07bcd98d766c875945b873f57e7880c4632cad9f1b3ca99944d7515261b720b",
    "140197bc205561987976170ecbe6adc03c92b2e73a0d028dad7f5799e20b840f",
    <snip>
    "34b8bc74cba0c32085ac73549fde2d9506fef0d3b192e143532776e1059ac799",
    "72548a9ecbda3261b35c103f3d6e86e1c941bcfaea12a6068201228c4cc34806"
  ],
  "mounts": [
    {
      "destination": "/etc/resolv.conf", "options": [ "rbind", "rshared", "rw" ],
      "source": "sandbox:///tmp/atlas/resolvconf/.+", "type": "bind"
    }
  ],
  "no_new_privileges": false,
  "seccomp_profile_sha256": "",
  "signals": [],
  "user": {
```

```
    "group_idnames": [ { "pattern": "", "strategy": "any" } ],
    "umask": "0022", "user_idname": { "pattern": "", "strategy": "any" }
  },
  "working_dir": "/"
},
{
  "allow_elevated": false,
  "allow_stdio_access": true,
  "capabilities": { <snip> },
  "command": [
    "/pause"
  ],
  "env_rules": [
    { "pattern": "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin", "required": false, "strategy": "string" },
    { "pattern": "TERM=xterm", "required": false, "strategy": "string" },
  ],
  "exec_processes": [],
  "layers": [
    "16b514057a06ad665f92c02863aca074fd5976c755d26bfff16365299169e8415"
  ],
  "mounts": [],
  "no_new_privileges": false,
  <snip>
  "working_dir": "/"
}
```

```
external_processes := [
  { "command": ["bash"], "env_rules": [{ "pattern": `PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin` } ]
}
```

```
allow_properties_access := true
allow_dump_stacks := true
allow_runtime_logging := true
allow_environment_variable_dropping := true
allow_unencrypted_scratch := false
allow_capability_dropping := true
```

```
mount_device := data.framework.mount_device
unmount_device := data.framework.unmount_device
mount_overlay := data.framework.mount_overlay
unmount_overlay := data.framework.unmount_overlay
create_container := data.framework.create_container
exec_in_container := data.framework.exec_in_container
exec_external := data.framework.exec_external
shutdown_container := data.framework.shutdown_container
signal_container_process := data.framework.signal_container_process
plan9_mount := data.framework.plan9_mount
plan9_unmount := data.framework.plan9_unmount
get_properties := data.framework.get_properties
dump_stacks := data.framework.dump_stacks
runtime_logging := data.framework.runtime_logging
load_fragment := data.framework.load_fragment
scratch_mount := data.framework.scratch_mount
scratch_unmount := data.framework.scratch_unmount
```

```
Reason := { "errors": data.framework.errors }
```

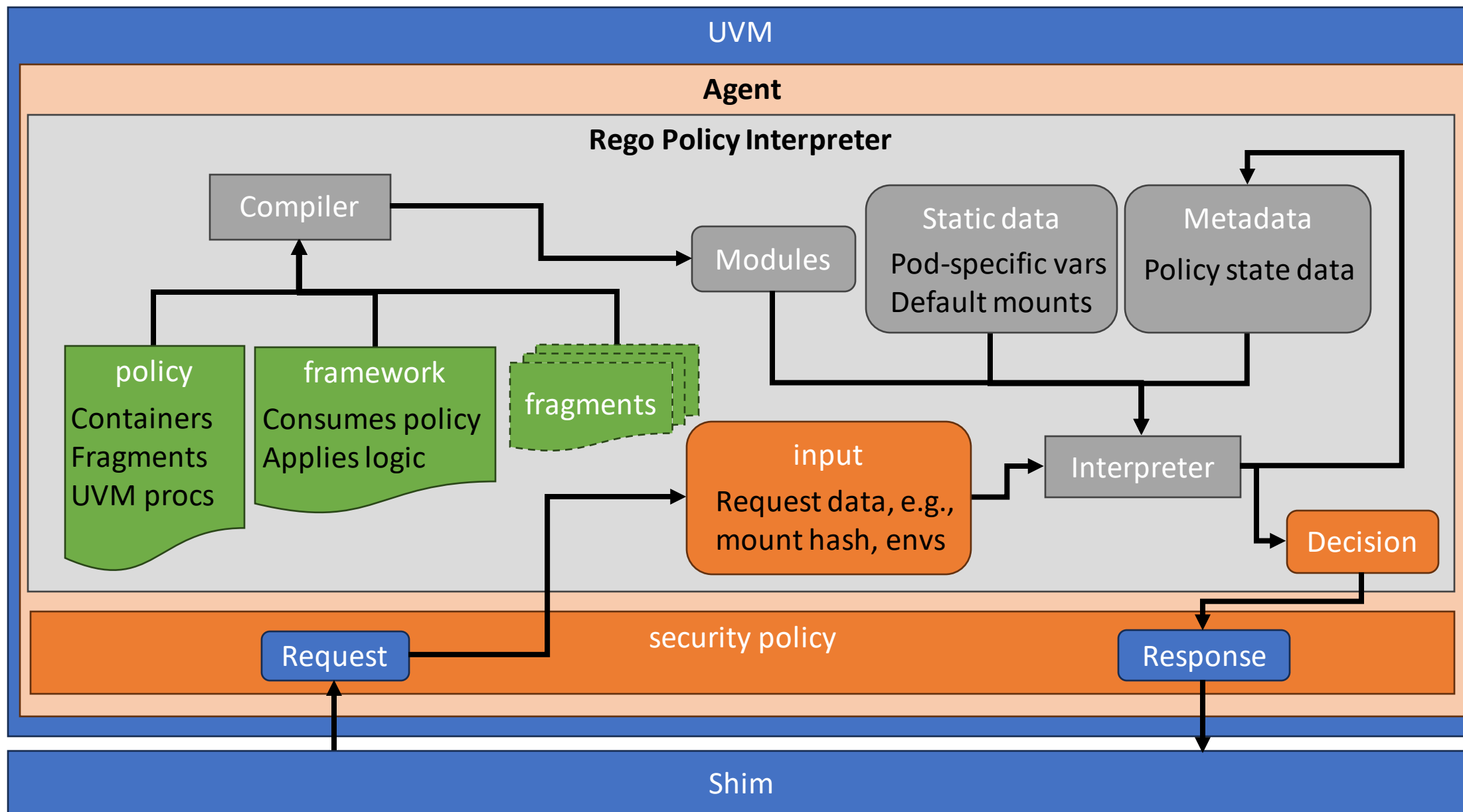

Execution Policy Engine in Rego

Framework code operates over that user policy

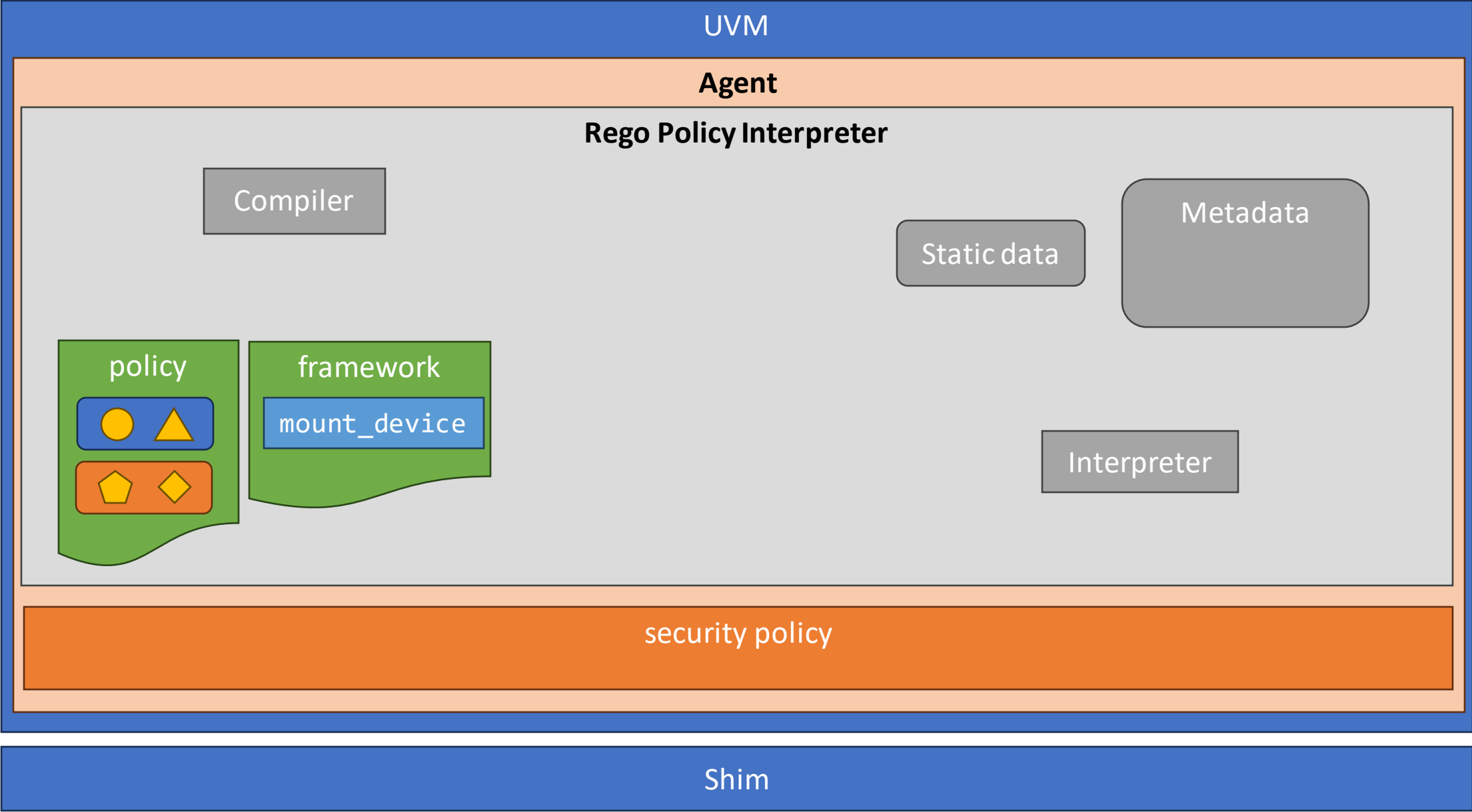
```
default mount_device := {"allowed": false}

mount_device := {"allowed": true} {
    some container in data.policy.containers
    some layer in container.layers
    input.deviceHash == layer
}
```

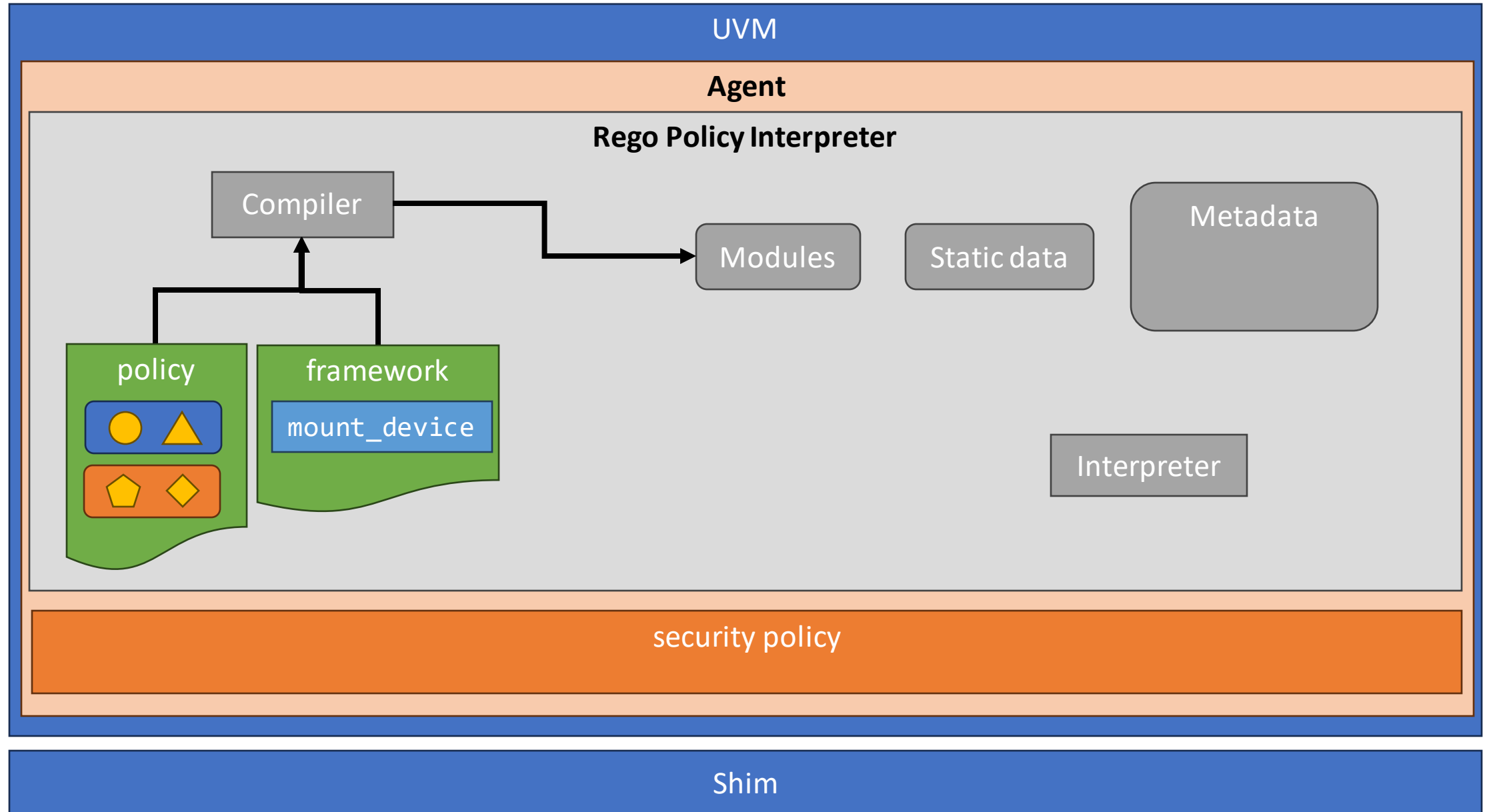
Confidential ACI – implementation overview



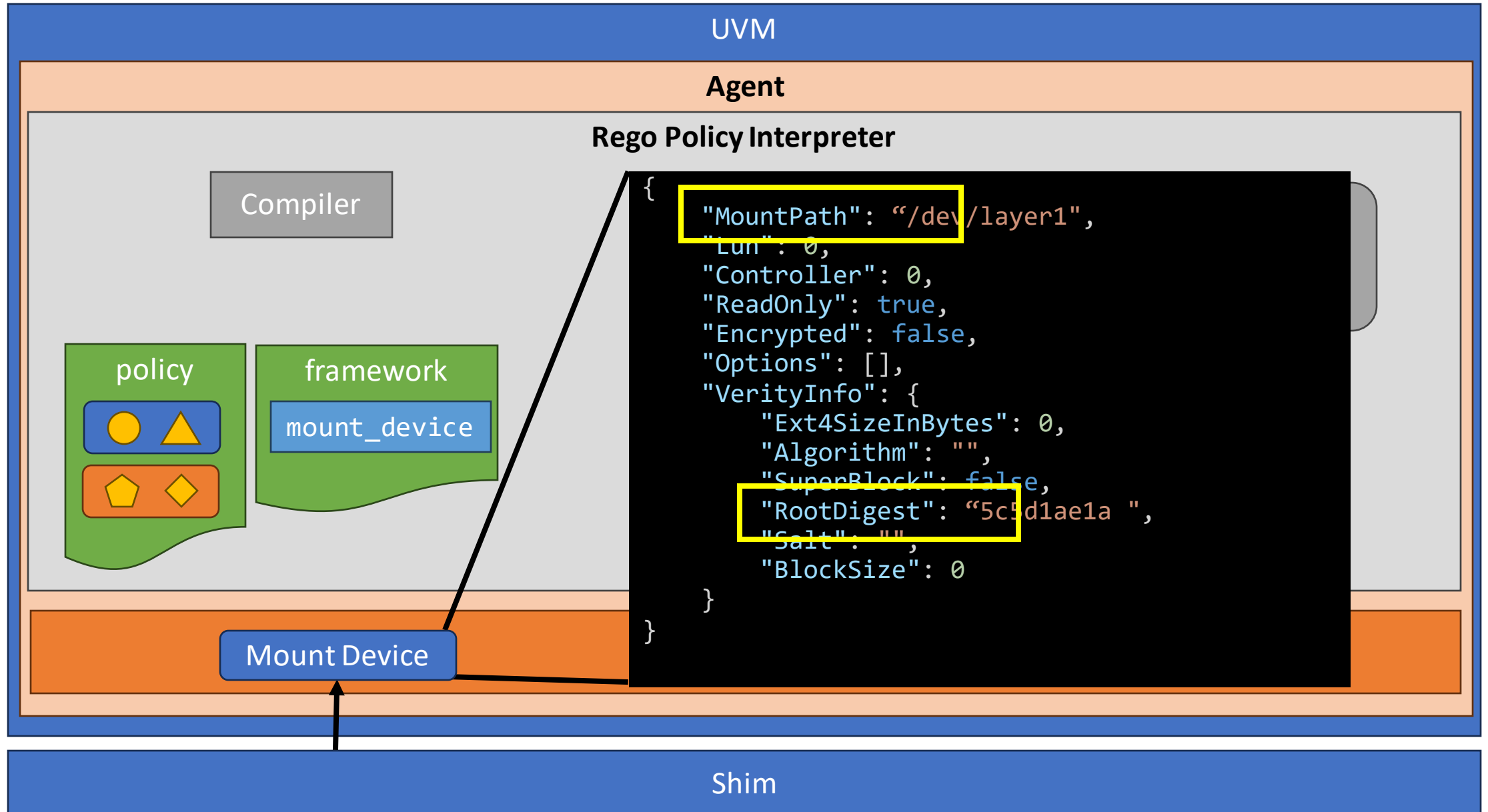
The shim provides the security policy, gcs provides the framework.



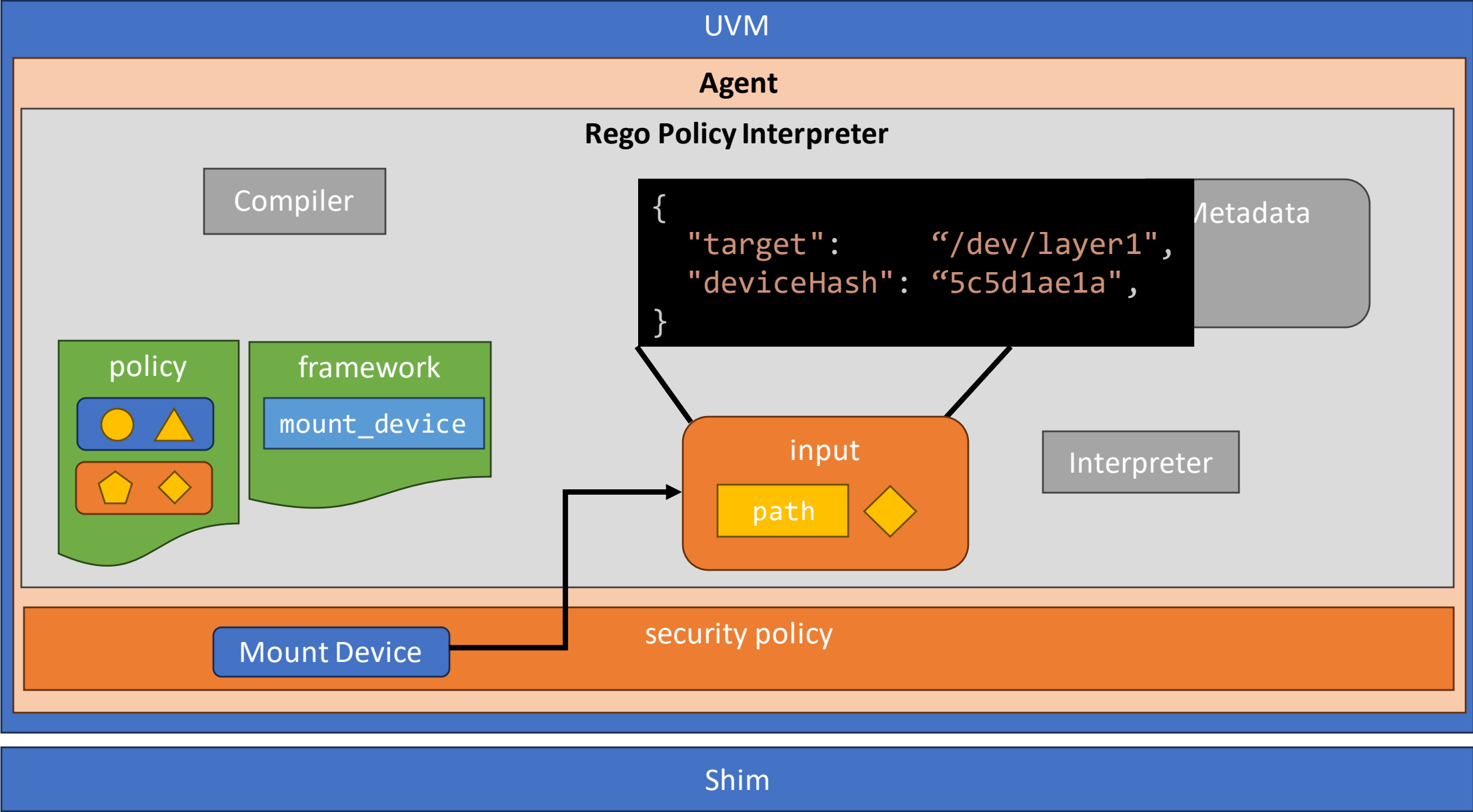
They get compiled into rego modules



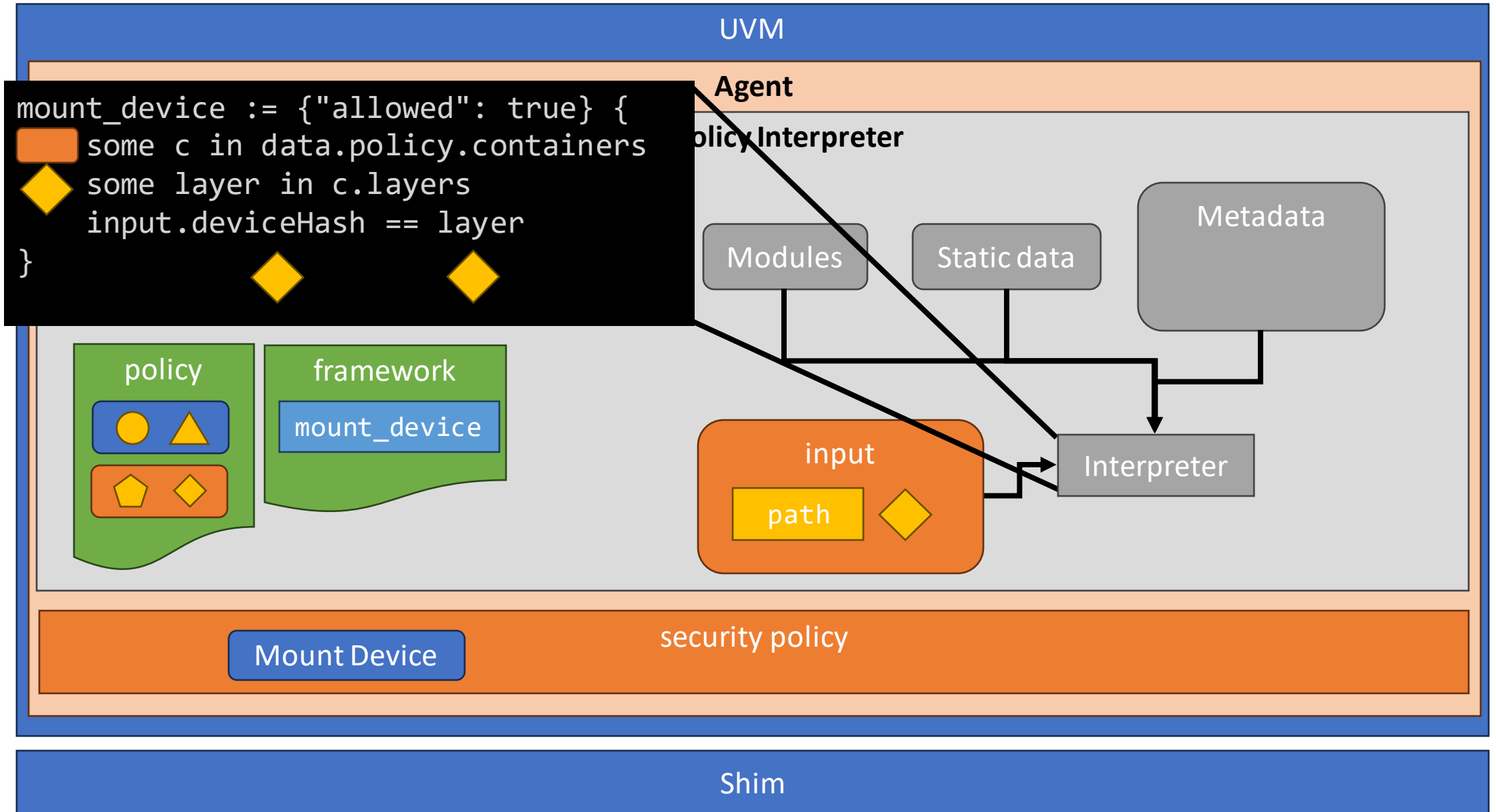
A request comes from the shim and is presented to the go security policy code



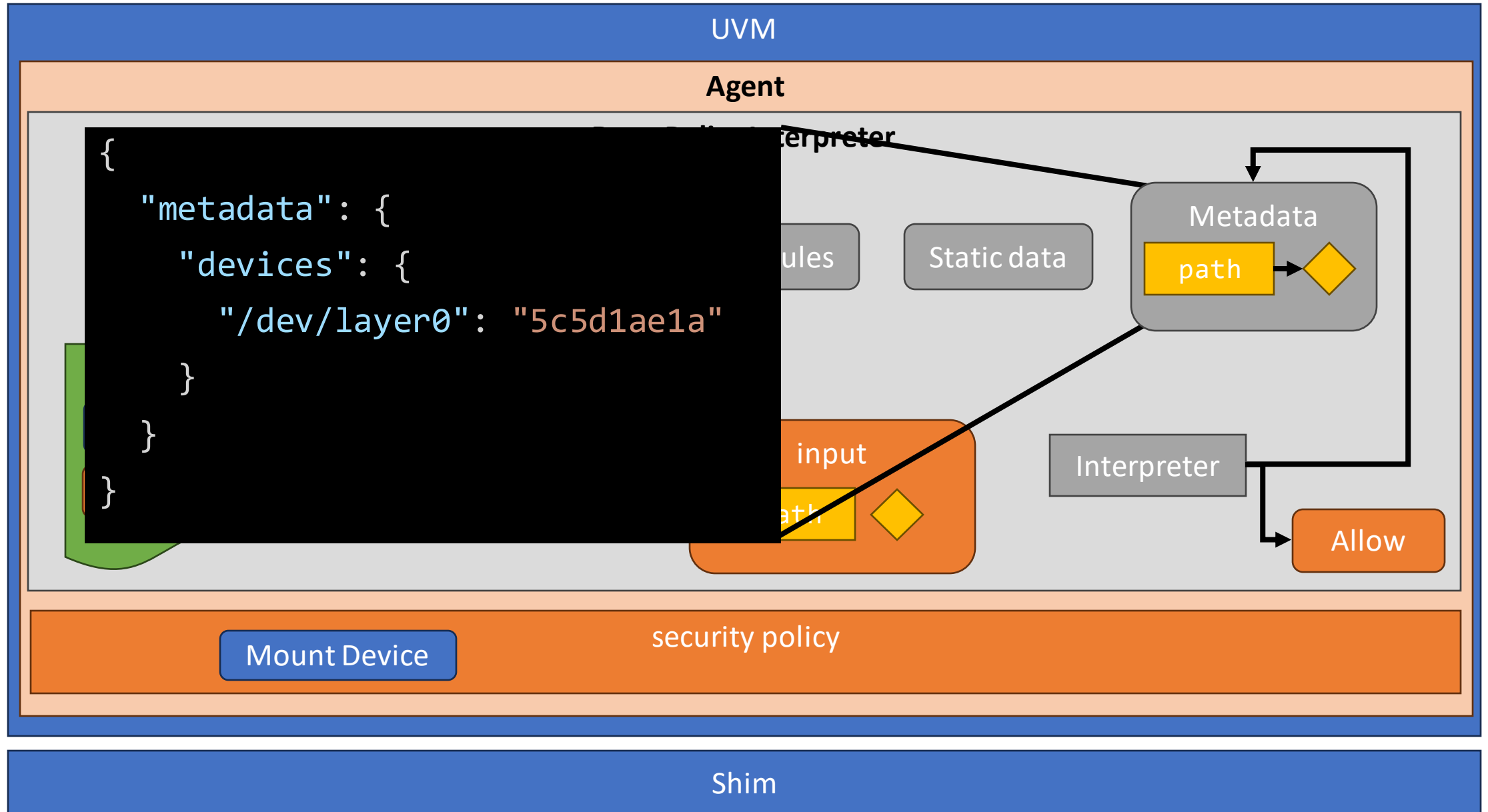
An “input” object is presented to the framework



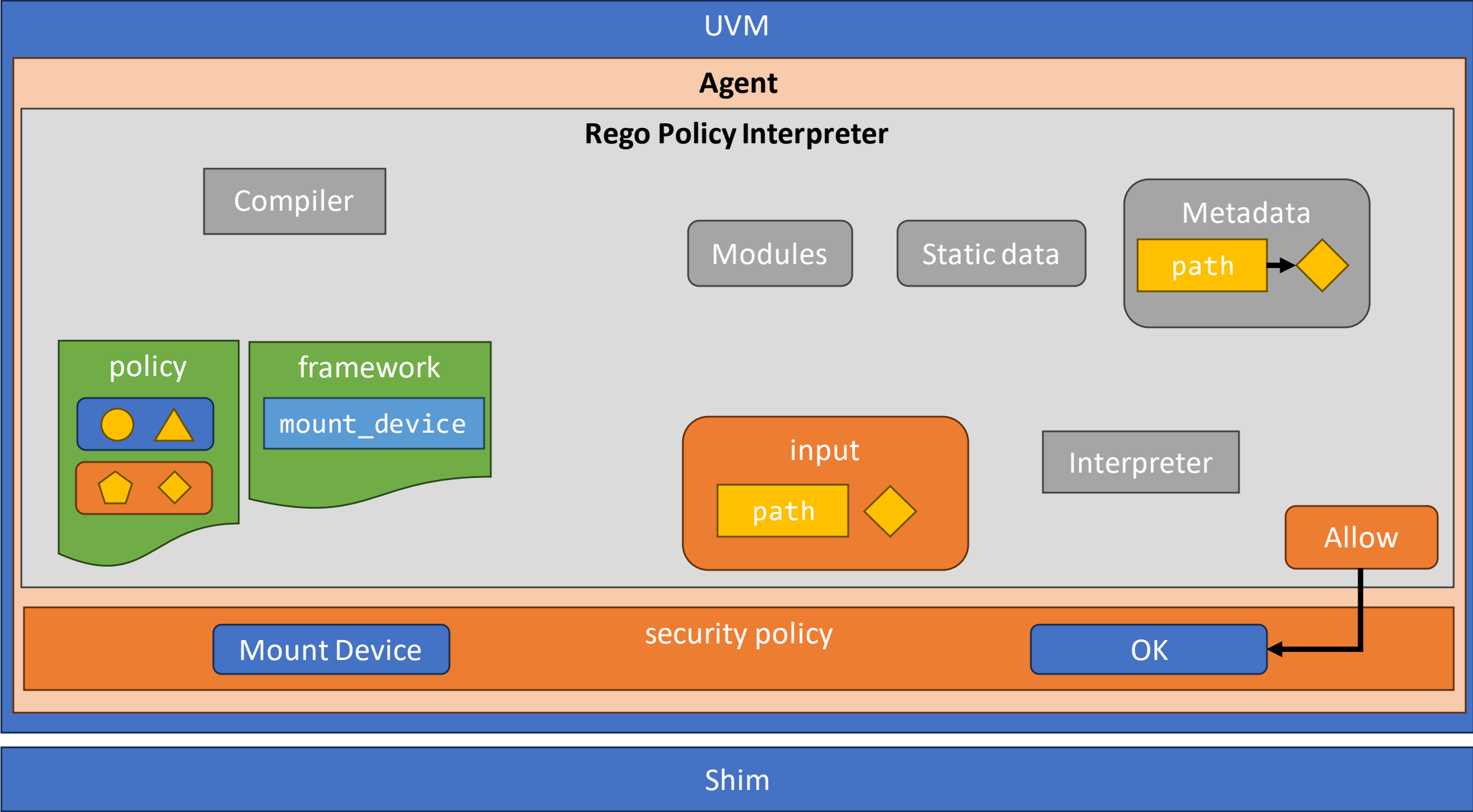
Rules in the framework test the request



If the test passes “metadata” state may be captured (narrowing etc)



If the test passed then “Allow” is returned to the go code



Managing change

A real cloud, real customers, no pontification and baseless assertions.

User containers change, infrastructure changes, Linux gets updated....

Change is everywhere and unavoidable.

Worse still, it is asynchronous, stuff goes forwards and backwards. How do we stop it going sideways?

We want the key release policy/relying party config to be stable in the face of approved change.

Fragments

- A piece of policy outside of the user supplied policy but allowed in by that policy
- A rule says “a fragment with an issuer DID of X (ie signed by a trusted party) and feed of Y
- Enables serviceability – this example allows ACI to update infra sidecars

```
fragments := [  
  {  
    "feed": "mcr.microsoft.com/aci/aci-cc-infra-fragment",  
    "includes": [  
      "containers",  
      "fragments"  
    ],  
    "issuer": "did:x509:0:sha256:I__iuL25oXEVFdTP_aBLx_eT1RPHbCQ_ECBQfYZpt9s::eku:1.3.6.1.4.1.311.76.59.1.3",  
    "minimum_svn": "1"  
  }  
]
```

Image Attached Fragments (real soon now...)

- A fragment which may come from the same container repository as a loaded container (using ORAS)
- Enables CUSTOMER serviceability – this example allows Parma to update interesting_app at will – so long as they can sign the fragment
- Enables stable policy (required for practical key management) in the face of realistic customer change, especially if they have thousands of containers

```
{
  "feed": "parma.azurecr.com/interesting_app",
  "includes": [
    "containers",
    "fragments"
  ],
  "issuer": "did:x509:0:sha256:0rdE7Z8Ayg1eUHa6zPutfJkft5-Y_8CIIZmlivElsb0::subject:CN:Test%20Leaf%20%28DO%20NOT%20TRUST%29",
  "minimum_svn": "42"
}
```

Standalone Fragments (future feature)

- A fragment which can come from the any container repository (using ORAS) – not specific to a container – path must be provided via arm template
- Allows customer to completely abstract away the policy
- The customer version of the ACI infra fragment

```
{
  "feed": "parma.azurecr.com/interesting_app",
  "includes": [
    "containers",
    "fragments"
  ],
  "issuer": "did:x509:0:sha256:0rdE7Z8Ayg1eUHa6zPutfJkft5-Y_8CIIZmlivElsb0::subject:CN:Test%20Leaf%20%28DO%20NOT%20TRUST%29",
  "minimum_svn": "42"
}
```

DID:x509 and COSE_Sign1

- [microsoft/did-x509: DRAFT: did:x509 Decentralized Identifier Method Specification \(github.com\)](https://github.com/microsoft/did-x509)
- <https://github.com/microsoft/didx509go>
- [CBOR Object Signing and Encryption \(COSE\) \(cose-wg.github.io\)](https://cose-wg.github.io)
- [microsoft/cosesign1go: A Go library to handle COSE Sign1 documents \(github.com\)](https://github.com/microsoft/cosesign1go)

Questions?

Spare/work in progress slides...

Trusting the platform

- The UVM is measured by the PSP into LAUNCH_MEASUREMENT
- A COSE Sign1 document, signed by Microsoft, asserts a given measurement is genuine.
- Given the UVM is genuine, we know it will properly enforce the supplied execution policy.
- A hash of the execution policy is bond to the hardware via the HOST_DATA field.
- The LAUNCH_MEASUREMENT and HOST_DATA are included in attestation reports generated by the PSP. That supported by a certificate chain from AMD.

}