Specification Version: *1.0.1*

Open Container Initiative Runtime Specification

The Open Container Initiative develops specifications for standards on Operating System process and application containers.

Abstract

The Open Container Initiative Runtime Specification aims to specify the configuration, execution environment, and lifecycle of a container.

A container's configuration is specified as the `config.json` for the supported platforms and details the fields that enable the creation of a container. The execution environment is specified to ensure that applications running inside a container have a consistent environment between runtimes along with common actions defined for the container's lifecycle.

Platforms

Platforms defined by this specification are:

- `linux`: runtime.md, config.md, features.md, config-linux.md, runtime-linux.md, and features-linux.md.
- `solaris`: runtime.md, config.md, features.md, and config-solaris.md.
- `windows`: runtime.md, config.md, features.md, and config-windows.md.
- `vm`: runtime.md, config.md, features.md, and config-vm.md.
- `zos`: runtime.md, config.md, features.md, and config-zos.md.

Table of Contents

  - Linux-specific Configuration
  - Solaris-specific Configuration
  - Windows-specific Configuration
  - Virtual-Machine-specific Configuration
  - z/OS-specific Configuration
- Features Structure
  - Linux-specific Features Structure
- Glossary

Notational Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119.

The key words "unspecified", "undefined", and "implementation-defined" are to be interpreted as described in the rationale for the C99 standard.

An implementation is not compliant for a given CPU architecture if it fails to satisfy one or more of the MUST, REQUIRED, or SHALL requirements for the platforms it implements. An implementation is compliant for a given CPU architecture if it satisfies all the MUST, REQUIRED, and SHALL requirements for the platforms it implements.

The 5 principles of Standard Containers

Define a unit of software delivery called a Standard Container. The goal of a Standard Container is to encapsulate a software component and all its dependencies in a format that is self-describing and portable, so that any compliant runtime can run it without extra dependencies, regardless of the underlying machine and the contents of the container.

The specification for Standard Containers defines:

1. configuration file formats
2. a set of standard operations
3. an execution environment.

A great analogy for this is the physical shipping container used by the transportation industry. Shipping containers are a fundamental unit of delivery, they can be lifted, stacked, locked, loaded, unloaded and labelled. Irrespective of their

contents, by standardizing the container itself it allowed for a consistent, more streamlined and efficient set of processes to be defined. For software Standard Containers offer similar functionality by being the fundamental, standardized, unit of delivery for a software package.

1. Standard operations

Standard Containers define a set of STANDARD OPERATIONS. They can be created, started, and stopped using standard container tools; copied and snapshotted using standard filesystem tools; and downloaded and uploaded using standard network tools.

2. Content-agnostic

Standard Containers are CONTENT-AGNOSTIC: all standard operations have the same effect regardless of the contents. They are started in the same way whether they contain a postgres database, a php application with its dependencies and application server, or Java build artifacts.

3. Infrastructure-agnostic

Standard Containers are INFRASTRUCTURE-AGNOSTIC: they can be run in any OCI supported infrastructure. For example, a standard container can be bundled on a laptop, uploaded to cloud storage, downloaded, run and snapshotted by a build server at a fiber hotel in Virginia, uploaded to 10 staging servers in a home-made private cloud cluster, then sent to 30 production instances across 3 public cloud regions.

4. Designed for automation

Standard Containers are DESIGNED FOR AUTOMATION: because they offer the same standard operations regardless of content and infrastructure, Standard Containers, are extremely well-suited for automation. In fact, you could say automation is their secret weapon.

Many things that once required time-consuming and error-prone human effort can now be programmed. Before Standard Containers, by the time a software component ran in production, it had been individually built, configured, bundled, documented, patched, vendored, templated, tweaked and instrumented by 10 different people on 10 different computers. Builds failed, libraries conflicted, mirrors crashed, post-it notes were lost, logs were misplaced, cluster updates were half-broken. The process was slow, inefficient and cost a fortune - and was entirely different depending on the language and infrastructure provider.

5. Industrial-grade delivery

Standard Containers make INDUSTRIAL-GRADE DELIVERY of software a reality. Leveraging all of the properties listed above, Standard Containers are enabling large and small enterprises to streamline and automate their software delivery pipelines. Whether it is in-house devOps flows, or external customer-based software delivery mechanisms, Standard Containers are changing the way the community thinks about software packaging and delivery.

Filesystem Bundle

Container Format

This section defines a format for encoding a container as a *filesystem bundle* - a set of files organized in a certain way, and containing all the necessary data and metadata for any compliant runtime to perform all standard operations against it. See also MacOS application bundles for a similar use of the term *bundle*.

The definition of a bundle is only concerned with how a container, and its configuration data, are stored on a local filesystem so that it can be consumed by a compliant runtime.

A Standard Container bundle contains all the information needed to load and run a container. This includes the following artifacts:

1. `config.json`: contains configuration data. This REQUIRED file MUST reside in the root of the bundle directory and MUST be named `config.json`. See `config.json` for more details.
2. container's root filesystem: the directory referenced by `root.path`, if that property is set in `config.json`.

When supplied, while these artifacts MUST all be present in a single directory on the local filesystem, that directory itself is not part of the bundle. In other words, a tar archive of a *bundle* will have these artifacts at the root of the archive, not nested within a top-level directory.

Runtime and Lifecycle

Scope of a Container

The entity using a runtime to create a container MUST be able to use the operations defined in this specification against that same container. Whether other entities using the same, or other, instance of the runtime can see that container is out of scope of this specification.

State

The state of a container includes the following properties:

- **`ociVersion`** (string, REQUIRED) is version of the Open Container Initiative Runtime Specification with which the state complies.
- **`id`** (string, REQUIRED) is the container's ID. This MUST be unique across all containers on this host. There is no requirement that it be unique across hosts.
- **`status`** (string, REQUIRED) is the runtime state of the container. The value MAY be one of:
  - `creating`: the container is being created (step 2 in the lifecycle)
  - `created`: the runtime has finished the create operation (after step 2 in the lifecycle), and the container process has neither exited nor executed the user-specified program
  - `running`: the container process has executed the user-specified program but has not exited (after step 8 in the lifecycle)
  - `stopped`: the container process has exited (step 10 in the lifecycle)

  Additional values MAY be defined by the runtime, however, they MUST be used to represent new runtime states not defined above.
- **`pid`** (int, REQUIRED when `status` is `created` or `running` on Linux, OPTIONAL on other platforms) is the ID of the container process. For hooks executed in the runtime namespace, it is the pid as seen by the runtime. For hooks executed in the container namespace, it is the pid as seen by the container.
- **`bundle`** (string, REQUIRED) is the absolute path to the container's bundle directory. This is provided so that consumers can find the container's configuration and root filesystem on the host.
- **`annotations`** (map, OPTIONAL) contains the list of annotations associated with the container. If no annotations were provided then this property MAY either be absent or an empty map.

The state MAY include additional properties.

When serialized in JSON, the format MUST adhere to the JSON Schema `schema/state-schema.json`.

See Query State for information on retrieving the state of a container.

Example

```
{
    "ociVersion": "0.2.0",
    "id": "oci-container1",
    "status": "running",
    "pid": 4422,
    "bundle": "/containers/redis",
    "annotations": {
        "myKey": "myValue"
    }
}
```

Lifecycle

The lifecycle describes the timeline of events that happen from when a container is created to when it ceases to exist.

1. OCI compliant runtime's `create` command is invoked with a reference to the location of the bundle and a unique identifier.
2. The container's runtime environment MUST be created according to the configuration in `config.json`. If the runtime is unable to create the environment specified in the `config.json`, it MUST generate an error. While the resources requested in the `config.json` MUST be created, the user-specified program (from `process`) MUST NOT be run at this time. Any updates to `config.json` after this step MUST NOT affect the container.
3. The `prestart` hooks MUST be invoked by the runtime. If any `prestart` hook fails, the runtime MUST generate an error, stop the container, and continue the lifecycle at step 12.
4. The `createRuntime` hooks MUST be invoked by the runtime. If any `createRuntime` hook fails, the runtime MUST generate an error, stop the container, and continue the lifecycle at step 12.
5. The `createContainer` hooks MUST be invoked by the runtime. If any `createContainer` hook fails, the runtime MUST generate an error, stop the container, and continue the lifecycle at step 12.
6. Runtime's `start` command is invoked with the unique identifier of the container.
7. The `startContainer` hooks MUST be invoked by the runtime. If any `startContainer` hook fails, the runtime MUST generate an error, stop the container, and continue the lifecycle at step 12.
8. The runtime MUST run the user-specified program, as specified by `process`.
9. The `poststart` hooks MUST be invoked by the runtime. If any `poststart` hook fails, the runtime MUST log a warning, but the remaining hooks and lifecycle continue as if the hook had succeeded.

10. The container process exits. This MAY happen due to erroring out, exiting, crashing or the runtime's `kill` operation being invoked.
11. Runtime's `delete` command is invoked with the unique identifier of the container.
12. The container MUST be destroyed by undoing the steps performed during create phase (step 2).
13. The `poststop` hooks MUST be invoked by the runtime. If any `poststop` hook fails, the runtime MUST log a warning, but the remaining hooks and lifecycle continue as if the hook had succeeded.

Errors

In cases where the specified operation generates an error, this specification does not mandate how, or even if, that error is returned or exposed to the user of an implementation. Unless otherwise stated, generating an error MUST leave the state of the environment as if the operation were never attempted - modulo any possible trivial ancillary changes such as logging.

Warnings

In cases where the specified operation logs a warning, this specification does not mandate how, or even if, that warning is returned or exposed to the user of an implementation. Unless otherwise stated, logging a warning does not change the flow of the operation; it MUST continue as if the warning had not been logged.

Operations

Unless otherwise stated, runtimes MUST support the following operations.

Note: these operations are not specifying any command-line APIs, and the parameters are inputs for general operations.

Query State

```
state <container-id>
```

This operation MUST generate an error if it is not provided the ID of a container. Attempting to query a container that does not exist MUST generate an error. This operation MUST return the state of a container as specified in the State section.

Create

`create <container-id> <path-to-bundle>`

This operation MUST generate an error if it is not provided a path to the bundle and the container ID to associate with the container. If the ID provided is not unique across all containers within the scope of the runtime, or is not valid in any other way, the implementation MUST generate an error and a new container MUST NOT be created. This operation MUST create a new container.

All of the properties configured in `config.json` except for `process` MUST be applied. `process.args` MUST NOT be applied until triggered by the `start` operation. The remaining `process` properties MAY be applied by this operation. If the runtime cannot apply a property as specified in the configuration, it MUST generate an error and a new container MUST NOT be created.

The runtime MAY validate `config.json` against this spec, either generically or with respect to the local system capabilities, before creating the container (step 2). Runtime callers who are interested in pre-create validation can run bundle-validation tools before invoking the create operation.

Any changes made to the `config.json` file after this operation will not have an effect on the container.

Start

`start <container-id>`

This operation MUST generate an error if it is not provided the container ID. Attempting to `start` a container that is not `created` MUST have no effect on the container and MUST generate an error. This operation MUST run the user-specified program as specified by `process`. This operation MUST generate an error if `process` was not set.

Kill

`kill <container-id> <signal>`

This operation MUST generate an error if it is not provided the container ID. Attempting to send a signal to a container that is neither `created` nor `running` MUST have no effect on the container and MUST generate an error. This operation MUST send the specified signal to the container process.

Delete

`delete <container-id>`

This operation MUST generate an error if it is not provided the container ID. Attempting to `delete` a container that is not `stopped` MUST have no effect on the container and MUST generate an error. Deleting a container MUST delete the resources that were created during the `create` step. Note that resources associated with the container, but not created by this container, MUST NOT be deleted. Once a container is deleted its ID MAY be used by a subsequent container.

Hooks

Many of the operations specified in this specification have "hooks" that allow for additional actions to be taken before or after each operation. See runtime configuration for hooks for more information.

Linux Runtime

File descriptors

By default, only the `stdin`, `stdout` and `stderr` file descriptors are kept open for the application by the runtime. The runtime MAY pass additional file descriptors to the application to support features such as socket activation. Some of the file descriptors MAY be redirected to `/dev/null` even though they are open.

Dev symbolic links

While creating the container (step 2 in the lifecycle), runtimes MUST create the following symlinks if the source file exists after processing `mounts`:

Source	Destination
/proc/self/fd	/dev/fd
/proc/self/fd/0	/dev/stdin
/proc/self/fd/1	/dev/stdout
/proc/self/fd/2	/dev/stderr

Configuration

This configuration file contains metadata necessary to implement standard operations against the container. This includes the process to run, environment variables to inject, sandboxing features to use, etc.

The canonical schema is defined in this document, but there is a JSON Schema in `schema/config-schema.json` and Go bindings in `specs-go/config.go`. Platform-specific configuration schema are defined in the platform-specific documents linked below. For properties that are only defined for some platforms, the Go property has a `platform` tag listing those protocols (e.g. `platform:"linux,solaris"`).

Below is a detailed description of each field defined in the configuration format and valid values are specified. Platform-specific fields are identified as such. For all platform-specific configuration values, the scope defined below in the Platform-specific configuration section applies.

Specification version

- **`ociVersion`** (string, REQUIRED) MUST be in SemVer v2.0.0 format and specifies the version of the Open Container Initiative Runtime Specification with which the bundle complies. The Open Container Initiative Runtime Specification follows semantic versioning and retains forward and backward compatibility within major versions. For example, if a configuration is compliant with version 1.1 of this specification, it is compatible with all runtimes that support any 1.1 or later release of this specification, but is not compatible with a runtime that supports 1.0 and not 1.1.

Example

```
"ociVersion": "0.1.0"
```

Root

`root` (object, OPTIONAL) specifies the container's root filesystem. On Windows, for Windows Server Containers, this field is REQUIRED. For Hyper-V Containers, this field MUST NOT be set.

On all other platforms, this field is REQUIRED.

- **`path`** (string, REQUIRED) Specifies the path to the root filesystem for the container.
  - On Windows, `path` MUST be a volume GUID path.
  - On POSIX platforms, `path` is either an absolute path or a relative path to the bundle. For example, with a bundle at `/to/bundle` and a root filesystem at `/to/bundle/rootfs`, the `path` value can be either `/to/bundle/rootfs` or `rootfs`. The value SHOULD be the conventional `rootfs`.

A directory MUST exist at the path declared by the field.

- **`readonly`** (bool, OPTIONAL) If true then the root filesystem MUST be read-only inside the container, defaults to false.
  - On Windows, this field MUST be omitted or false.

Example (POSIX platforms)

```
"root": {
    "path": "rootfs",
    "readonly": true
}
```

Example (Windows)

```
"root": {
    "path": "\\\\?\\Volume{ec84d99e-3f02-11e7-ac6c-00155d7682cf}\\"
}
```

Mounts

`mounts` (array of objects, OPTIONAL) specifies additional mounts beyond **`root`**. The runtime MUST mount entries in the listed order. For Linux, the parameters are as documented in mount(2) system call man page. For Solaris, the mount entry corresponds to the 'fs' resource in the zonecfg(1M) man page.

- **`destination`** (string, REQUIRED) Destination of mount point: path inside container.
  - Linux: This value SHOULD be an absolute path. For compatibility with old tools and configurations, it MAY be a relative path, in which case it MUST be interpreted as relative to "/". Relative paths are **deprecated**.
  - Windows: This value MUST be an absolute path. One mount destination MUST NOT be nested within another mount (e.g., c:\foo and c:\foo\bar).
  - Solaris: This value MUST be an absolute path. Corresponds to "dir" of the fs resource in zonecfg(1M).
  - For all other platforms: This value MUST be an absolute path.
- **`source`** (string, OPTIONAL) A device name, but can also be a file or directory name for bind mounts or a dummy. Path values for bind mounts are either absolute or relative to the bundle. A mount is a bind mount if it has either **`bind`** or **`rbind`** in the options.

  - Windows: a local directory on the filesystem of the container host. UNC paths and mapped drives are not supported.
  - Solaris: corresponds to "special" of the fs resource in zonecfg(1M).

- `options` (array of strings, OPTIONAL) Mount options of the filesystem to be used.
  - Linux: See Linux mount options below.
  - Solaris: corresponds to "options" of the fs resource in zonecfg(1M).
  - Windows: runtimes MUST support `ro`, mounting the filesystem read-only when `ro` is given.

Linux mount options

Runtimes MUST/SHOULD/MAY implement the following option strings for Linux:

Option name	Requirement	Description
`async`	MUST	[^1]
`atime`	MUST	[^1]
`bind`	MUST	Bind mount [^2]
`defaults`	MUST	[^1]
`dev`	MUST	[^1]
`diratime`	MUST	[^1]
`dirsync`	MUST	[^1]
`exec`	MUST	[^1]
`iversion`	MUST	[^1]
`lazytime`	MUST	[^1]
`loud`	MUST	[^1]
`mand`	MAY	[^1] (Deprecated in kernel 5.15, util-linux 2.38)
`noatime`	MUST	[^1]
`nodev`	MUST	[^1]
`nodiratime`	MUST	[^1]
`noexec`	MUST	[^1]
`noiversion`	MUST	[^1]
`nolazytime`	MUST	[^1]
`nomand`	MAY	[^1]
`norelatime`	MUST	[^1]
`nostrictatime`	MUST	[^1]
`nosuid`	MUST	[^1]
`nosymfollow`	SHOULD	[^1] (Introduced in kernel 5.10, util-linux 2.38)
`private`	MUST	Bind mount propagation [^2]
`ratime`	SHOULD	Recursive `atime` [^3]
`rbind`	MUST	Recursive bind mount [^2]

Option name	Requirement	Description
`rdev`	SHOULD	Recursive `dev` [^3]
`rdiratime`	SHOULD	Recursive `diratime` [^3]
`relatime`	MUST	[^1]
`remount`	MUST	[^1]
`rexec`	SHOULD	Recursive `dev` [^3]
`rnoatime`	SHOULD	Recursive `noatime` [^3]
`rnodiratime`	SHOULD	Recursive `nodiratime` [^3]
`rnoexec`	SHOULD	Recursive `noexec` [^3]
`rnorelatime`	SHOULD	Recursive `norelatime` [^3]
`rnostrictatime`	SHOULD	Recursive `nostrictatime` [^3]
`rnosuid`	SHOULD	Recursive `nosuid` [^3]
`rnosymfollow`	SHOULD	Recursive `nosymfollow` [^3]
`ro`	MUST	[^1]
`rprivate`	MUST	Bind mount propagation [^2]
`rrelatime`	SHOULD	Recursive `relatime` [^3]
`rro`	SHOULD	Recursive `ro` [^3]
`rrw`	SHOULD	Recursive `rw` [^3]
`rshared`	MUST	Bind mount propagation [^2]
`rslave`	MUST	Bind mount propagation [^2]
`rstrictatime`	SHOULD	Recursive `strictatime` [^3]
`rsuid`	SHOULD	Recursive `suid` [^3]
`rsymfollow`	SHOULD	Recursive `symfollow` [^3]
`runbindable`	MUST	Bind mount propagation [^2]
`rw`	MUST	[^1]
`shared`	MUST	[^1]
`silent`	MUST	[^1]
`slave`	MUST	Bind mount propagation [^2]
`strictatime`	MUST	[^1]
`suid`	MUST	[^1]
`symfollow`	SHOULD	Opposite of `nosymfollow`
`sync`	MUST	[^1]
`tmpcopyup`	MAY	copy up the contents to a tmpfs
`unbindable`	MUST	Bind mount propagation [^2]
`idmap`	SHOULD	Indicates that the mount has `uidMappings` and `gidMappings` specified. This option SHOULD NOT be passed to the underlying `mount(2)` call. If supported, the runtime MUST return an error if this option is provided and either of `uidMappings` or `gidMappings` are empty or not present.

Option name	Requirement	Description
`ridmap`	SHOULD	Indicates that the mount has `uidMappings` and `gidMappings` specified, and the mapping is applied recursively [^3]. This option SHOULD NOT be passed to the underlying `mount(2)` call. If supported, the runtime MUST return an error if this option is provided and either of `uidMappings` or `gidMappings` are empty or not present.

[^1]: Corresponds to `mount(8)` (filesystem-independent). [^2]: Corresponds to bind mounts and shared subtrees. [^3]: These `AT_RECURSIVE` options need kernel 5.12 or later. See `mount_setattr(2)`

The "MUST" options correspond to `mount(8)`.

Runtimes MAY also implement custom option strings that are not listed in the table above. If a custom option string is already recognized by `mount(8)`, the runtime SHOULD follow the behavior of `mount(8)`.

Runtimes SHOULD treat unknown options as filesystem-specific ones) and pass those as a comma-separated string to the fifth (`const void *data`) argument of `mount(2)`.

Example (Windows)

```
"mounts": [
    {
        "destination": "C:\\folder-inside-container",
        "source": "C:\\folder-on-host",
        "options": ["ro"]
    }
]
```

POSIX-platform Mounts

For POSIX platforms the `mounts` structure has the following fields:

- **`type`** (string, OPTIONAL) The type of the filesystem to be mounted.
  - Linux: filesystem types supported by the kernel as listed in */proc/filesystems* (e.g., "minix", "ext2", "ext3", "jfs", "xfs", "reiserfs", "msdos", "proc", "nfs", "iso9660"). For bind mounts (when `options` include either `bind` or `rbind`), the type is a dummy, often "none" (not listed in */proc/filesystems*).

  - Solaris: corresponds to "type" of the fs resource in zonecfg(1M).

- **uidMappings** (array of type LinuxIDMapping, OPTIONAL) The mapping to convert UIDs from the source file system to the destination mount point. This SHOULD be implemented using `mount_setattr(MOUNT_ATTR_IDMAP)`, available since Linux 5.12. If specified, the **options** field of the **mounts** structure SHOULD contain either **idmap** or **ridmap** to specify whether the mapping should be applied recursively for **rbind** mounts, as well as to ensure that older runtimes will not silently ignore this field. The format is the same as user namespace mappings. If specified, it MUST be specified along with **gidMappings**.
- **gidMappings** (array of type LinuxIDMapping, OPTIONAL) The mapping to convert GIDs from the source file system to the destination mount point. This SHOULD be implemented using `mount_setattr(MOUNT_ATTR_IDMAP)`, available since Linux 5.12. If specified, the **options** field of the **mounts** structure SHOULD contain either **idmap** or **ridmap** to specify whether the mapping should be applied recursively for **rbind** mounts, as well as to ensure that older runtimes will not silently ignore this field. For more details see **uidMappings**. If specified, it MUST be specified along with **uidMappings**.

Example (Linux)

```
"mounts": [
    {
        "destination": "/tmp",
        "type": "tmpfs",
        "source": "tmpfs",
        "options": ["nosuid","strictatime","mode=755","size=65536k"]
    },
    {
        "destination": "/data",
        "type": "none",
        "source": "/volumes/testing",
        "options": ["rbind","rw"]
    }
]
```

Example (Solaris)

```
"mounts": [
    {
        "destination": "/opt/local",
        "type": "lofs",
        "source": "/usr/local",
```

```
        "options": ["ro","nodevices"]
    },
    {
        "destination": "/opt/sfw",
        "type": "lofs",
        "source": "/opt/sfw"
    }
]
```

Process

process (object, OPTIONAL) specifies the container process. This property is REQUIRED when **start** is called.

- **terminal** (bool, OPTIONAL) specifies whether a terminal is attached to the process, defaults to false. As an example, if set to true on Linux a pseudoterminal pair is allocated for the process and the pseudoterminal pty is duplicated on the process's standard streams.
- **consoleSize** (object, OPTIONAL) specifies the console size in characters of the terminal. Runtimes MUST ignore **consoleSize** if **terminal** is **false** or unset.
  - **height** (uint, REQUIRED)
  - **width** (uint, REQUIRED)
- **cwd** (string, REQUIRED) is the working directory that will be set for the executable. This value MUST be an absolute path.
- **env** (array of strings, OPTIONAL) with the same semantics as IEEE Std 1003.1-2008's **environ**.
- **args** (array of strings, OPTIONAL) with similar semantics to IEEE Std 1003.1-2008 **execvp**'s *argv*. This specification extends the IEEE standard in that at least one entry is REQUIRED (non-Windows), and that entry is used with the same semantics as **execvp**'s *file*. This field is OPTIONAL on Windows, and **commandLine** is REQUIRED if this field is omitted.
- **commandLine** (string, OPTIONAL) specifies the full command line to be executed on Windows. This is the preferred means of supplying the command line on Windows. If omitted, the runtime will fall back to escaping and concatenating fields from **args** before making the system call into Windows.

POSIX process

For systems that support POSIX rlimits (for example Linux and Solaris), the **process** object supports the following process-specific properties:

- `rlimits` (array of objects, OPTIONAL) allows setting resource limits for the process. Each entry has the following structure:
  - `type` (string, REQUIRED) the platform resource being limited.
    * Linux: valid values are defined in the `getrlimit(2)` man page, such as `RLIMIT_MSGQUEUE`.
    * Solaris: valid values are defined in the `getrlimit(3)` man page, such as `RLIMIT_CORE`.

    The runtime MUST generate an error for any values which cannot be mapped to a relevant kernel interface. For each entry in `rlimits`, a `getrlimit(3)` on `type` MUST succeed. For the following properties, `rlim` refers to the status returned by the `getrlimit(3)` call.
  - `soft` (uint64, REQUIRED) the value of the limit enforced for the corresponding resource. `rlim.rlim_cur` MUST match the configured value.
  - `hard` (uint64, REQUIRED) the ceiling for the soft limit that could be set by an unprivileged process. `rlim.rlim_max` MUST match the configured value. Only a privileged process (e.g. one with the `CAP_SYS_RESOURCE` capability) can raise a hard limit.

  If `rlimits` contains duplicated entries with same `type`, the runtime MUST generate an error.

Linux Process

For Linux-based systems, the `process` object supports the following process-specific properties.

- `apparmorProfile` (string, OPTIONAL) specifies the name of the AppArmor profile for the process. For more information about AppArmor, see AppArmor documentation.
- `capabilities` (object, OPTIONAL) is an object containing arrays that specifies the sets of capabilities for the process. Valid values are defined in the capabilities(7) man page, such as `CAP_CHOWN`. Any value which cannot be mapped to a relevant kernel interface, or cannot be granted otherwise MUST be logged as a warning by the runtime. Runtimes SHOULD NOT fail if the container configuration requests capabilities that cannot be granted, for example, if the runtime operates in a restricted environment with a limited set of capabilities. `capabilities` contains the following properties:
  - `effective` (array of strings, OPTIONAL) the `effective` field is an array of effective capabilities that are kept for the process.

  - `bounding` (array of strings, OPTIONAL) the `bounding` field is an array of bounding capabilities that are kept for the process.
  - `inheritable` (array of strings, OPTIONAL) the `inheritable` field is an array of inheritable capabilities that are kept for the process.
  - `permitted` (array of strings, OPTIONAL) the `permitted` field is an array of permitted capabilities that are kept for the process.
  - `ambient` (array of strings, OPTIONAL) the `ambient` field is an array of ambient capabilities that are kept for the process.

- `noNewPrivileges` (bool, OPTIONAL) setting `noNewPrivileges` to true prevents the process from gaining additional privileges. As an example, the `no_new_privs` article in the kernel documentation has information on how this is achieved using a `prctl` system call on Linux.

- `oomScoreAdj` *(int, OPTIONAL)* adjusts the oom-killer score in `[pid]/oom_score_adj` for the process's `[pid]` in a proc pseudo-filesystem. If `oomScoreAdj` is set, the runtime MUST set `oom_score_adj` to the given value. If `oomScoreAdj` is not set, the runtime MUST NOT change the value of `oom_score_adj`.

  This is a per-process setting, where as `disableOOMKiller` is scoped for a memory cgroup. For more information on how these two settings work together, see the memory cgroup documentation section 10. OOM Contol.

- `scheduler` (object, OPTIONAL) is an object describing the scheduler properties for the process. The `scheduler` contains the following properties:

  - `policy` (string, REQUIRED) represents the scheduling policy. A valid list of values is:

    * `SCHED_OTHER`
    * `SCHED_FIFO`
    * `SCHED_RR`
    * `SCHED_BATCH`
    * `SCHED_ISO`
    * `SCHED_IDLE`
    * `SCHED_DEADLINE`

  - `nice` (int32, OPTIONAL) is the nice value for the process, affecting its priority. A lower nice value corresponds to a higher priority. If not set, the runtime must use the value 0.

  - `priority` (int32, OPTIONAL) represents the static priority of the process, used by real-time policies like SCHED_FIFO and SCHED_RR. If not set, the runtime must use the value 0.

  - `flags` (array of strings, OPTIONAL) is an array of strings representing scheduling flags. A valid list of values is:

    * `SCHED_FLAG_RESET_ON_FORK`

    * SCHED_FLAG_RECLAIM
    * SCHED_FLAG_DL_OVERRUN
    * SCHED_FLAG_KEEP_POLICY
    * SCHED_FLAG_KEEP_PARAMS
    * SCHED_FLAG_UTIL_CLAMP_MIN
    * SCHED_FLAG_UTIL_CLAMP_MAX
  - **runtime** (uint64, OPTIONAL) represents the amount of time in nanoseconds during which the process is allowed to run in a given period, used by the deadline scheduler. If not set, the runtime must use the value 0.
  - **deadline** (uint64, OPTIONAL) represents the absolute deadline for the process to complete its execution, used by the deadline scheduler. If not set, the runtime must use the value 0.
  - **period** (uint64, OPTIONAL) represents the length of the period in nanoseconds used for determining the process runtime, used by the deadline scheduler. If not set, the runtime must use the value 0.

- **selinuxLabel** (string, OPTIONAL) specifies the SELinux label for the process. For more information about SELinux, see SELinux documentation.

- **ioPriority** (object, OPTIONAL) configures the I/O priority settings for the container's processes within the process group. The I/O priority settings will be automatically applied to the entire process group, affecting all processes within the container. The following properties are available:

  - **class** (string, REQUIRED) specifies the I/O scheduling class. Possible values are IOPRIO_CLASS_RT, IOPRIO_CLASS_BE, and IOPRIO_CLASS_IDLE.
  - **priority** (int, REQUIRED) specifies the priority level within the class. The value should be an integer ranging from 0 (highest) to 7 (lowest).

User

The user for the process is a platform-specific structure that allows specific control over which user the process runs as.

POSIX-platform User For POSIX platforms the **user** structure has the following fields:

- **uid** (int, REQUIRED) specifies the user ID in the container namespace.
- **gid** (int, REQUIRED) specifies the group ID in the container namespace.

- **umask** (int, OPTIONAL) specifies the [umask][umask_2] of the user. If unspecified, the umask should not be changed from the calling process' umask.
- **additionalGids** (array of ints, OPTIONAL) specifies additional group IDs in the container namespace to be added to the process.

Note: symbolic name for uid and gid, such as uname and gname respectively, are left to upper levels to derive (i.e. `/etc/passwd` parsing, NSS, etc)

Example (Linux)

```
"process": {
    "terminal": true,
    "consoleSize": {
        "height": 25,
        "width": 80
    },
    "user": {
        "uid": 1,
        "gid": 1,
        "umask": 63,
        "additionalGids": [5, 6]
    },
    "env": [
        "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin",
        "TERM=xterm"
    ],
    "cwd": "/root",
    "args": [
        "sh"
    ],
    "apparmorProfile": "acme_secure_profile",
    "selinuxLabel": "system_u:system_r:svirt_lxc_net_t:s0:c124,c675",
    "ioPriority": {
        "class": "IOPRIO_CLASS_IDLE",
        "priority": 4
    },
    "noNewPrivileges": true,
    "capabilities": {
        "bounding": [
            "CAP_AUDIT_WRITE",
            "CAP_KILL",
            "CAP_NET_BIND_SERVICE"
        ],
       "permitted": [
```

```
                "CAP_AUDIT_WRITE",
                "CAP_KILL",
                "CAP_NET_BIND_SERVICE"
            ],
           "inheritable": [
                "CAP_AUDIT_WRITE",
                "CAP_KILL",
                "CAP_NET_BIND_SERVICE"
            ],
            "effective": [
                "CAP_AUDIT_WRITE",
                "CAP_KILL"
            ],
            "ambient": [
                "CAP_NET_BIND_SERVICE"
            ]
        },
        "rlimits": [
            {
                "type": "RLIMIT_NOFILE",
                "hard": 1024,
                "soft": 1024
            }
        ]
}
```

Example (Solaris)

```
"process": {
    "terminal": true,
    "consoleSize": {
        "height": 25,
        "width": 80
    },
    "user": {
        "uid": 1,
        "gid": 1,
        "umask": 7,
        "additionalGids": [2, 8]
    },
    "env": [
        "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin",
        "TERM=xterm"
    ],
    "cwd": "/root",
```

```
    "args": [
        "/usr/bin/bash"
    ]
}
```

Windows User For Windows based systems the user structure has the following fields:

- **username** (string, OPTIONAL) specifies the user name for the process.

Example (Windows)

```
"process": {
    "terminal": true,
    "user": {
        "username": "containeradministrator"
    },
    "env": [
        "VARIABLE=1"
    ],
    "cwd": "c:\\foo",
    "args": [
        "someapp.exe",
    ]
}
```

Hostname

- **hostname** (string, OPTIONAL) specifies the container's hostname as seen by processes running inside the container. On Linux, for example, this will change the hostname in the container UTS namespace. Depending on your namespace configuration, the container UTS namespace may be the runtime UTS namespace.

Example

```
"hostname": "mrsdalloway"
```

Domainname

- **domainname** (string, OPTIONAL) specifies the container's domainname as seen by processes running inside the container. On Linux, for example, this

will change the domainname in the container UTS namespace. Depending on your namespace configuration, the container UTS namespace may be the runtime UTS namespace.

Example

```
"domainname": "foobarbaz.test"
```

Platform-specific configuration

- **`linux`** (object, OPTIONAL) Linux-specific configuration. This MAY be set if the target platform of this spec is **`linux`**.
- **`windows`** (object, OPTIONAL) Windows-specific configuration. This MUST be set if the target platform of this spec is **`windows`**.
- **`solaris`** (object, OPTIONAL) Solaris-specific configuration. This MAY be set if the target platform of this spec is **`solaris`**.
- **`vm`** (object, OPTIONAL) Virtual-machine-specific configuration. This MAY be set if the target platform and architecture of this spec support hardware virtualization.
- **`zos`** (object, OPTIONAL) z/OS-specific configuration. This MAY be set if the target platform of this spec is **`zos`**.

Example (Linux)

```
{
    "linux": {
        "namespaces": [
            {
                "type": "pid"
            }
        ]
    }
}
```

POSIX-platform Hooks

For POSIX platforms, the configuration structure supports **`hooks`** for configuring custom actions related to the lifecycle of the container.

- **`hooks`** (object, OPTIONAL) MAY contain any of the following properties:
  - **`prestart`** (array of objects, OPTIONAL, **DEPRECATED**) is an array of **`prestart`** hooks.

    * Entries in the array contain the following properties:
        · **`path`** (string, REQUIRED) with similar semantics to IEEE Std 1003.1-2008 `execv`'s *path*. This specification extends the IEEE standard in that **`path`** MUST be absolute.
        · **`args`** (array of strings, OPTIONAL) with the same semantics as IEEE Std 1003.1-2008 `execv`'s *argv*.
        · **`env`** (array of strings, OPTIONAL) with the same semantics as IEEE Std 1003.1-2008's `environ`.
        · **`timeout`** (int, OPTIONAL) is the number of seconds before aborting the hook. If set, `timeout` MUST be greater than zero.
    * The value of `path` MUST resolve in the runtime namespace.
    * The `prestart` hooks MUST be executed in the runtime namespace.
- **`createRuntime`** (array of objects, OPTIONAL) is an array of `createRuntime` hooks.
    * Entries in the array contain the following properties (the entries are identical to the entries in the deprecated `prestart` hooks):
        · **`path`** (string, REQUIRED) with similar semantics to IEEE Std 1003.1-2008 `execv`'s *path*. This specification extends the IEEE standard in that **`path`** MUST be absolute.
        · **`args`** (array of strings, OPTIONAL) with the same semantics as IEEE Std 1003.1-2008 `execv`'s *argv*.
        · **`env`** (array of strings, OPTIONAL) with the same semantics as IEEE Std 1003.1-2008's `environ`.
        · **`timeout`** (int, OPTIONAL) is the number of seconds before aborting the hook. If set, `timeout` MUST be greater than zero.
    * The value of `path` MUST resolve in the runtime namespace.
    * The `createRuntime` hooks MUST be executed in the runtime namespace.
- **`createContainer`** (array of objects, OPTIONAL) is an array of `createContainer` hooks.
    * Entries in the array have the same schema as `createRuntime` entries.
    * The value of `path` MUST resolve in the runtime namespace.
    * The `createContainer` hooks MUST be executed in the container namespace.
- **`startContainer`** (array of objects, OPTIONAL) is an array of `startContainer` hooks.
    * Entries in the array have the same schema as `createRuntime` entries.
    * The value of `path` MUST resolve in the container namespace.

    - The `startContainer` hooks MUST be executed in the container namespace.
- `poststart` (array of objects, OPTIONAL) is an array of `poststart` hooks.
    - Entries in the array have the same schema as `createRuntime` entries.
    - The value of `path` MUST resolve in the runtime namespace.
    - The `poststart` hooks MUST be executed in the runtime namespace.
- `poststop` (array of objects, OPTIONAL) is an array of `poststop` hooks.
    - Entries in the array have the same schema as `createRuntime` entries.
    - The value of `path` MUST resolve in the runtime namespace.
    - The `poststop` hooks MUST be executed in the runtime namespace.

Hooks allow users to specify programs to run before or after various lifecycle events. Hooks MUST be called in the listed order. The state of the container MUST be passed to hooks over stdin so that they may do work appropriate to the current state of the container.

Prestart

The `prestart` hooks MUST be called as part of the `create` operation after the runtime environment has been created (according to the configuration in config.json) but before the `pivot_root` or any equivalent operation has been executed. On Linux, for example, they are called after the container namespaces are created, so they provide an opportunity to customize the container (e.g. the network namespace could be specified in this hook). The `prestart` hooks MUST be called before the `createRuntime` hooks.

Note: `prestart` hooks were deprecated in favor of `createRuntime`, `createContainer` and `startContainer` hooks, which allow more granular hook control during the create and start phase.

The `prestart` hooks' path MUST resolve in the runtime namespace. The `prestart` hooks MUST be executed in the runtime namespace.

CreateRuntime Hooks

The `createRuntime` hooks MUST be called as part of the `create` operation after the runtime environment has been created (according to the configuration in config.json) but before the `pivot_root` or any equivalent operation has been executed.

The `createRuntime` hooks' path MUST resolve in the runtime namespace. The `createRuntime` hooks MUST be executed in the runtime namespace.

On Linux, for example, they are called after the container namespaces are created, so they provide an opportunity to customize the container (e.g. the network namespace could be specified in this hook).

The definition of `createRuntime` hooks is currently underspecified and hooks authors, should only expect from the runtime that the mount namespace have been created and the mount operations performed. Other operations such as cgroups and SELinux/AppArmor labels might not have been performed by the runtime.

CreateContainer Hooks

The `createContainer` hooks MUST be called as part of the `create` operation after the runtime environment has been created (according to the configuration in config.json) but before the `pivot_root` or any equivalent operation has been executed. The `createContainer` hooks MUST be called after the `createRuntime` hooks.

The `createContainer` hooks' path MUST resolve in the runtime namespace. The `createContainer` hooks MUST be executed in the container namespace.

For example, on Linux this would happen before the `pivot_root` operation is executed but after the mount namespace was created and setup.

The definition of `createContainer` hooks is currently underspecified and hooks authors, should only expect from the runtime that the mount namespace and different mounts will be setup. Other operations such as cgroups and SELinux/AppArmor labels might not have been performed by the runtime.

StartContainer Hooks

The `startContainer` hooks MUST be called before the user-specified process is executed as part of the `start` operation. This hook can be used to execute some operations in the container, for example running the `ldconfig` binary on linux before the container process is spawned.

The `startContainer` hooks' path MUST resolve in the container namespace. The `startContainer` hooks MUST be executed in the container namespace.

Poststart

The `poststart` hooks MUST be called after the user-specified process is executed but before the `start` operation returns. For example, this hook can notify the user that the container process is spawned.

The `poststart` hooks' path MUST resolve in the runtime namespace. The `poststart` hooks MUST be executed in the runtime namespace.

Poststop

The `poststop` hooks MUST be called after the container is deleted but before the `delete` operation returns. Cleanup or debugging functions are examples of such a hook.

The `poststop` hooks' path MUST resolve in the runtime namespace. The `poststop` hooks MUST be executed in the runtime namespace.

Summary

See the below table for a summary of hooks and when they are called:

Name	Namespace	When
`prestart` (Deprecated)	runtime	After the start operation is called but before the user-specified program command is executed.
`createRuntime`	runtime	During the create operation, after the runtime environment has been created and before the pivot root or any equivalent operation.
`createContainer`	container	During the create operation, after the runtime environment has been created and before the pivot root or any equivalent operation.
`startContainer`	container	After the start operation is called but before the user-specified program command is executed.
`poststart`	runtime	After the user-specified process is executed but before the start operation returns.
`poststop`	runtime	After the container is deleted but before the delete operation returns.

Example

```
"hooks": {
    "prestart": [
        {
            "path": "/usr/bin/fix-mounts",
            "args": ["fix-mounts", "arg1", "arg2"],
            "env":  [ "key1=value1"]
        },
        {
            "path": "/usr/bin/setup-network"
```

```
        }
    ],
    "createRuntime": [
        {
            "path": "/usr/bin/fix-mounts",
            "args": ["fix-mounts", "arg1", "arg2"],
            "env":  [ "key1=value1"]
        },
        {
            "path": "/usr/bin/setup-network"
        }
    ],
    "createContainer": [
        {
            "path": "/usr/bin/mount-hook",
            "args": ["-mount", "arg1", "arg2"],
            "env":  [ "key1=value1"]
        }
    ],
    "startContainer": [
        {
            "path": "/usr/bin/refresh-ldcache"
        }
    ],
    "poststart": [
        {
            "path": "/usr/bin/notify-start",
            "timeout": 5
        }
    ],
    "poststop": [
        {
            "path": "/usr/sbin/cleanup.sh",
            "args": ["cleanup.sh", "-f"]
        }
    ]
}
```

Annotations

`annotations` (object, OPTIONAL) contains arbitrary metadata for the container. This information MAY be structured or unstructured. Annotations MUST be a key-value map. If there are no annotations then this property MAY either be absent or an empty map.

Keys MUST be strings. Keys MUST NOT be an empty string. Keys SHOULD be named using a reverse domain notation - e.g. `com.example.myKey`. Keys using the `org.opencontainers` namespace are reserved and MUST NOT be used by subsequent specifications. Runtimes MUST handle unknown annotation keys like any other unknown property.

Values MUST be strings. Values MAY be an empty string.

```
"annotations": {
    "com.example.gpu-cores": "2"
}
```

Extensibility

Runtimes MAY log unknown properties but MUST otherwise ignore them. That includes not generating errors if they encounter an unknown property.

Valid values

Runtimes MUST generate an error when invalid or unsupported values are encountered. Unless support for a valid value is explicitly required, runtimes MAY choose which subset of the valid values it will support.

Configuration Schema Example

Here is a full example `config.json` for reference.

```
{
    "ociVersion": "1.0.1",
    "process": {
        "terminal": true,
        "user": {
            "uid": 1,
            "gid": 1,
            "additionalGids": [
                5,
                6
            ]
        },
        "args": [
            "sh"
        ],
        "env": [
```

```
        "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin",
        "TERM=xterm"
    ],
    "cwd": "/",
    "capabilities": {
        "bounding": [
            "CAP_AUDIT_WRITE",
            "CAP_KILL",
            "CAP_NET_BIND_SERVICE"
        ],
        "permitted": [
            "CAP_AUDIT_WRITE",
            "CAP_KILL",
            "CAP_NET_BIND_SERVICE"
        ],
        "inheritable": [
            "CAP_AUDIT_WRITE",
            "CAP_KILL",
            "CAP_NET_BIND_SERVICE"
        ],
        "effective": [
            "CAP_AUDIT_WRITE",
            "CAP_KILL"
        ],
        "ambient": [
            "CAP_NET_BIND_SERVICE"
        ]
    },
    "rlimits": [
        {
            "type": "RLIMIT_CORE",
            "hard": 1024,
            "soft": 1024
        },
        {
            "type": "RLIMIT_NOFILE",
            "hard": 1024,
            "soft": 1024
        }
    ],
    "apparmorProfile": "acme_secure_profile",
    "oomScoreAdj": 100,
    "selinuxLabel": "system_u:system_r:svirt_lxc_net_t:s0:c124,c675",
    "ioPriority": {
        "class": "IOPRIO_CLASS_IDLE",
        "priority": 4
```

```
    },
    "noNewPrivileges": true
},
"root": {
    "path": "rootfs",
    "readonly": true
},
"hostname": "slartibartfast",
"mounts": [
    {
        "destination": "/proc",
        "type": "proc",
        "source": "proc"
    },
    {
        "destination": "/dev",
        "type": "tmpfs",
        "source": "tmpfs",
        "options": [
            "nosuid",
            "strictatime",
            "mode=755",
            "size=65536k"
        ]
    },
    {
        "destination": "/dev/pts",
        "type": "devpts",
        "source": "devpts",
        "options": [
            "nosuid",
            "noexec",
            "newinstance",
            "ptmxmode=0666",
            "mode=0620",
            "gid=5"
        ]
    },
    {
        "destination": "/dev/shm",
        "type": "tmpfs",
        "source": "shm",
        "options": [
            "nosuid",
            "noexec",
            "nodev",
```

```
                "mode=1777",
                "size=65536k"
            ]
        },
        {
            "destination": "/dev/mqueue",
            "type": "mqueue",
            "source": "mqueue",
            "options": [
                "nosuid",
                "noexec",
                "nodev"
            ]
        },
        {
            "destination": "/sys",
            "type": "sysfs",
            "source": "sysfs",
            "options": [
                "nosuid",
                "noexec",
                "nodev"
            ]
        },
        {
            "destination": "/sys/fs/cgroup",
            "type": "cgroup",
            "source": "cgroup",
            "options": [
                "nosuid",
                "noexec",
                "nodev",
                "relatime",
                "ro"
            ]
        }
    ],
    "hooks": {
        "prestart": [
            {
                "path": "/usr/bin/fix-mounts",
                "args": [
                    "fix-mounts",
                    "arg1",
                    "arg2"
                ],
```

```
            "env": [
                "key1=value1"
            ]
        },
        {
            "path": "/usr/bin/setup-network"
        }
    ],
    "poststart": [
        {
            "path": "/usr/bin/notify-start",
            "timeout": 5
        }
    ],
    "poststop": [
        {
            "path": "/usr/sbin/cleanup.sh",
            "args": [
                "cleanup.sh",
                "-f"
            ]
        }
    ]
},
"linux": {
    "devices": [
        {
            "path": "/dev/fuse",
            "type": "c",
            "major": 10,
            "minor": 229,
            "fileMode": 438,
            "uid": 0,
            "gid": 0
        },
        {
            "path": "/dev/sda",
            "type": "b",
            "major": 8,
            "minor": 0,
            "fileMode": 432,
            "uid": 0,
            "gid": 0
        }
    ],
    "uidMappings": [
```

```
        {
            "containerID": 0,
            "hostID": 1000,
            "size": 32000
        }
    ],
    "gidMappings": [
        {
            "containerID": 0,
            "hostID": 1000,
            "size": 32000
        }
    ],
    "sysctl": {
        "net.ipv4.ip_forward": "1",
        "net.core.somaxconn": "256"
    },
    "cgroupsPath": "/myRuntime/myContainer",
    "resources": {
        "network": {
            "classID": 1048577,
            "priorities": [
                {
                    "name": "eth0",
                    "priority": 500
                },
                {
                    "name": "eth1",
                    "priority": 1000
                }
            ]
        },
        "pids": {
            "limit": 32771
        },
        "hugepageLimits": [
            {
                "pageSize": "2MB",
                "limit": 9223372036854772000
            },
            {
                "pageSize": "64KB",
                "limit": 1000000
            }
        ],
        "memory": {
```

```
        "limit": 536870912,
        "reservation": 536870912,
        "swap": 536870912,
        "kernel": -1,
        "kernelTCP": -1,
        "swappiness": 0,
        "disableOOMKiller": false
    },
    "cpu": {
        "shares": 1024,
        "quota": 1000000,
        "period": 500000,
        "realtimeRuntime": 950000,
        "realtimePeriod": 1000000,
        "cpus": "2-3",
        "idle": 1,
        "mems": "0-7"
    },
    "devices": [
        {
            "allow": false,
            "access": "rwm"
        },
        {
            "allow": true,
            "type": "c",
            "major": 10,
            "minor": 229,
            "access": "rw"
        },
        {
            "allow": true,
            "type": "b",
            "major": 8,
            "minor": 0,
            "access": "r"
        }
    ],
    "blockIO": {
        "weight": 10,
        "leafWeight": 10,
        "weightDevice": [
            {
                "major": 8,
                "minor": 0,
                "weight": 500,
```

```
                    "leafWeight": 300
                },
                {
                    "major": 8,
                    "minor": 16,
                    "weight": 500
                }
            ],
            "throttleReadBpsDevice": [
                {
                    "major": 8,
                    "minor": 0,
                    "rate": 600
                }
            ],
            "throttleWriteIOPSDevice": [
                {
                    "major": 8,
                    "minor": 16,
                    "rate": 300
                }
            ]
        }
    },
    "rootfsPropagation": "slave",
    "seccomp": {
        "defaultAction": "SCMP_ACT_ALLOW",
        "architectures": [
            "SCMP_ARCH_X86",
            "SCMP_ARCH_X32"
        ],
        "syscalls": [
            {
                "names": [
                    "getcwd",
                    "chmod"
                ],
                "action": "SCMP_ACT_ERRNO"
            }
        ]
    },
    "timeOffsets": {
        "monotonic": {
            "secs": 172800,
            "nanosecs": 0
        },
```

```
        "boottime": {
            "secs": 604800,
            "nanosecs": 0
        }
    },
    "namespaces": [
        {
            "type": "pid"
        },
        {
            "type": "network"
        },
        {
            "type": "ipc"
        },
        {
            "type": "uts"
        },
        {
            "type": "mount"
        },
        {
            "type": "user"
        },
        {
            "type": "cgroup"
        },
        {
            "type": "time"
        }
    ],
    "maskedPaths": [
        "/proc/kcore",
        "/proc/latency_stats",
        "/proc/timer_stats",
        "/proc/sched_debug"
    ],
    "readonlyPaths": [
        "/proc/asound",
        "/proc/bus",
        "/proc/fs",
        "/proc/irq",
        "/proc/sys",
        "/proc/sysrq-trigger"
    ],
    "mountLabel": "system_u:object_r:svirt_sandbox_file_t:s0:c715,c811"
```

```
    },
    "annotations": {
        "com.example.key1": "value1",
        "com.example.key2": "value2"
    }
}
```

Linux Container Configuration

This document describes the schema for the Linux-specific section of the container configuration. The Linux container specification uses various kernel features like namespaces, cgroups, capabilities, LSM, and filesystem jails to fulfill the spec.

Default Filesystems

The Linux ABI includes both syscalls and several special file paths. Applications expecting a Linux environment will very likely expect these file paths to be set up correctly.

The following filesystems SHOULD be made available in each container's filesystem:

Path	Type
/proc	proc
/sys	sysfs
/dev/pts	devpts
/dev/shm	tmpfs

Namespaces

A namespace wraps a global system resource in an abstraction that makes it appear to the processes within the namespace that they have their own isolated instance of the global resource. Changes to the global resource are visible to other processes that are members of the namespace, but are invisible to other processes. For more information, see the namespaces(7) man page.

Namespaces are specified as an array of entries inside the `namespaces` root field. The following parameters can be specified to set up namespaces:

- **`type`** *(string, REQUIRED)* - namespace type. The following namespace types SHOULD be supported:

  - **pid** processes inside the container will only be able to see other processes inside the same container or inside the same pid namespace.
  - **network** the container will have its own network stack.
  - **mount** the container will have an isolated mount table.
  - **ipc** processes inside the container will only be able to communicate to other processes inside the same container via system level IPC.
  - **uts** the container will be able to have its own hostname and domain name.
  - **user** the container will be able to remap user and group IDs from the host to local users and groups within the container.
  - **cgroup** the container will have an isolated view of the cgroup hierarchy.
  - **time** the container will be able to have its own clocks.

- **path** *(string, OPTIONAL)* - namespace file. This value MUST be an absolute path in the runtime mount namespace. The runtime MUST place the container process in the namespace associated with that **path**. The runtime MUST generate an error if **path** is not associated with a namespace of type **type**.

  If **path** is not specified, the runtime MUST create a new container namespace of type **type**.

If a namespace type is not specified in the **namespaces** array, the container MUST inherit the runtime namespace of that type. If a **namespaces** field contains duplicated namespaces with same **type**, the runtime MUST generate an error.

Example

```
"namespaces": [
    {
        "type": "pid",
        "path": "/proc/1234/ns/pid"
    },
    {
        "type": "network",
        "path": "/var/run/netns/neta"
    },
    {
        "type": "mount"
    },
    {
        "type": "ipc"
    },
    {
```

```
        "type": "uts"
    },
    {
        "type": "user"
    },
    {
        "type": "cgroup"
    },
    {
        "type": "time"
    }
]
```

User namespace mappings

uidMappings (array of objects, OPTIONAL) describes the user namespace uid mappings from the host to the container. **gidMappings** (array of objects, OPTIONAL) describes the user namespace gid mappings from the host to the container.

Each entry has the following structure:

- **containerID** *(uint32, REQUIRED)* - is the starting uid/gid in the container.
- **hostID** *(uint32, REQUIRED)* - is the starting uid/gid on the host to be mapped to *containerID*.
- **size** *(uint32, REQUIRED)* - is the number of ids to be mapped.

The runtime SHOULD NOT modify the ownership of referenced filesystems to realize the mapping. Note that the number of mapping entries MAY be limited by the kernel.

Example

```
"uidMappings": [
    {
        "containerID": 0,
        "hostID": 1000,
        "size": 32000
    }
],
"gidMappings": [
    {
        "containerID": 0,
```

```
        "hostID": 1000,
        "size": 32000
    }
]
```

Offset for Time Namespace

timeOffsets (object, OPTIONAL) sets the offset for Time Namespace. For more information see the time_namespaces.

The name of the clock is the entry key. Entry values are objects with the following properties:

- **secs** *(int64, OPTIONAL)* - is the offset of clock (in seconds) in the container.
- **nanosecs** *(uint32, OPTIONAL)* - is the offset of clock (in nanoseconds) in the container.

Devices

devices (array of objects, OPTIONAL) lists devices that MUST be available in the container. The runtime MAY supply them however it likes (with `mknod`, by bind mounting from the runtime mount namespace, using symlinks, etc.).

Each entry has the following structure:

- **type** *(string, REQUIRED)* - type of device: `c`, `b`, `u` or `p`. More info in mknod(1).
- **path** *(string, REQUIRED)* - full path to device inside container. If a file already exists at `path` that does not match the requested device, the runtime MUST generate an error. The path MAY be anywhere in the container filesystem, notably outside of `/dev`.
- **major, minor** *(int64, REQUIRED unless `type` is `p`)* - major, minor numbers for the device.
- **fileMode** *(uint32, OPTIONAL)* - file mode for the device. You can also control access to devices with cgroups.
- **uid** *(uint32, OPTIONAL)* - id of device owner in the container namespace.
- **gid** *(uint32, OPTIONAL)* - id of device group in the container namespace.

The same `type`, `major` and `minor` SHOULD NOT be used for multiple devices.

Containers MAY NOT access any device node that is not either explicitly referenced in the **devices** array or listed as being part of the default devices. Rationale: runtimes based on virtual machines need to be able to adjust the node devices, and accessing device nodes that were not adjusted could have undefined behaviour.

Example

```
"devices": [
    {
        "path": "/dev/fuse",
        "type": "c",
        "major": 10,
        "minor": 229,
        "fileMode": 438,
        "uid": 0,
        "gid": 0
    },
    {
        "path": "/dev/sda",
        "type": "b",
        "major": 8,
        "minor": 0,
        "fileMode": 432,
        "uid": 0,
        "gid": 0
    }
]
```

Default Devices

In addition to any devices configured with this setting, the runtime MUST also supply:

- `/dev/null`
- `/dev/zero`
- `/dev/full`
- `/dev/random`
- `/dev/urandom`
- `/dev/tty`
- `/dev/console` is set up if `terminal` is enabled in the config by bind mounting the pseudoterminal pty to `/dev/console`.
- `/dev/ptmx`. A bind-mount or symlink of the container's `/dev/pts/ptmx`.

Control groups

Also known as cgroups, they are used to restrict resource usage for a container and handle device access. cgroups provide controls (through controllers) to restrict cpu, memory, IO, pids, network and RDMA resources for the container. For more information, see the kernel cgroups documentation.

A runtime MAY, during a particular container operation, such as create, start, or exec, check if the container cgroup is fit for purpose, and MUST generate an error if such a check fails. For example, a frozen cgroup or (for create operation) a non-empty cgroup. The reason for this is that accepting such configurations could cause container operation outcomes that users may not anticipate or understand, such as operation on one container inadvertently affecting other containers.

Cgroups Path

`cgroupsPath` (string, OPTIONAL) path to the cgroups. It can be used to either control the cgroups hierarchy for containers or to run a new process in an existing container.

The value of `cgroupsPath` MUST be either an absolute path or a relative path.

- In the case of an absolute path (starting with `/`), the runtime MUST take the path to be relative to the cgroups mount point.
- In the case of a relative path (not starting with `/`), the runtime MAY interpret the path relative to a runtime-determined location in the cgroups hierarchy.

If the value is specified, the runtime MUST consistently attach to the same place in the cgroups hierarchy given the same value of `cgroupsPath`. If the value is not specified, the runtime MAY define the default cgroups path. Runtimes MAY consider certain `cgroupsPath` values to be invalid, and MUST generate an error if this is the case.

Implementations of the Spec can choose to name cgroups in any manner. The Spec does not include naming schema for cgroups. The Spec does not support per-controller paths for the reasons discussed in the cgroupv2 documentation. The cgroups will be created if they don't exist.

You can configure a container's cgroups via the `resources` field of the Linux configuration. Do not specify `resources` unless limits have to be updated. For example, to run a new process in an existing container without updating limits, `resources` need not be specified.

Runtimes MAY attach the container process to additional cgroup controllers beyond those necessary to fulfill the `resources` settings.

Cgroup ownership

Runtimes MAY, according to the following rules, change (or cause to be changed) the owner of the container's cgroup to the host uid that maps to the

value of `process.user.uid` in the container namespace; that is, the user that will execute the container process.

Runtimes SHOULD NOT change the ownership of container cgroups when cgroups v1 is in use. Cgroup delegation is not secure in cgroups v1.

A runtime SHOULD NOT change the ownership of a container cgroup unless it will also create a new cgroup namespace for the container. Typically this occurs when the `linux.namespaces` array contains an object with `type` equal to `"cgroup"` and `path` unset.

Runtimes SHOULD change the cgroup ownership if and only if the cgroup filesystem is to be mounted read/write; that is, when the configuration's `mounts` array contains an object where:

- The `source` field is equal to `"cgroup"`
- The `destination` field is equal to `"/sys/fs/cgroup"`
- The `options` field does not contain the value `"ro"`

If the configuration does not specify such a mount, the runtime SHOULD NOT change the cgroup ownership.

A runtime that changes the cgroup ownership SHOULD only change the ownership of the container's cgroup directory and files within that directory that are listed in `/sys/kernel/cgroup/delegate`. See `cgroups(7)` for details about this file. Note that not all files listed in `/sys/kernel/cgroup/delegate` necessarily exist in every cgroup. Runtimes MUST NOT fail in this scenario, and SHOULD change the ownership of the listed files that do exist in the cgroup.

If the `/sys/kernel/cgroup/delegate` file does not exist, the runtime MUST fall back to using the following list of files:

```
cgroup.procs
cgroup.subtree_control
cgroup.threads
```

The runtime SHOULD NOT change the ownership of any other files. Changing other files may allow the container to elevate its own resource limits or perform other unwanted behaviour.

Example

```
"cgroupsPath": "/myRuntime/myContainer",
"resources": {
    "memory": {
    "limit": 100000,
```

```
    "reservation": 200000
    },
    "devices": [
        {
            "allow": false,
            "access": "rwm"
        }
    ]
}
```

Allowed Device list

`devices` (array of objects, OPTIONAL) configures the allowed device list. The runtime MUST apply entries in the listed order.

Each entry has the following structure:

- **`allow`** *(boolean, REQUIRED)* - whether the entry is allowed or denied.
- **`type`** *(string, OPTIONAL)* - type of device: `a` (all), `c` (char), or `b` (block). Unset values mean "all", mapping to `a`.
- **`major, minor`** *(int64, OPTIONAL)* - major, minor numbers for the device. Unset values mean "all", mapping to `*` in the filesystem API.
- **`access`** *(string, OPTIONAL)* - cgroup permissions for device. A composition of `r` (read), `w` (write), and `m` (mknod).

Example

```
"devices": [
    {
        "allow": false,
        "access": "rwm"
    },
    {
        "allow": true,
        "type": "c",
        "major": 10,
        "minor": 229,
        "access": "rw"
    },
    {
        "allow": true,
        "type": "b",
        "major": 8,
        "minor": 0,
        "access": "r"
```

```
    }
]
```

Memory

`memory` (object, OPTIONAL) represents the cgroup subsystem `memory` and it's used to set limits on the container's memory usage. For more information, see the kernel cgroups documentation about memory.

Values for memory specify the limit in bytes, or `-1` for unlimited memory.

- **`limit`** *(int64, OPTIONAL)* - sets limit of memory usage
- **`reservation`** *(int64, OPTIONAL)* - sets soft limit of memory usage
- **`swap`** *(int64, OPTIONAL)* - sets limit of memory+Swap usage
- **`kernel`** *(int64, OPTIONAL, NOT RECOMMENDED)* - sets hard limit for kernel memory
- **`kernelTCP`** *(int64, OPTIONAL, NOT RECOMMENDED)* - sets hard limit for kernel TCP buffer memory

The following properties do not specify memory limits, but are covered by the `memory` controller:

- **`swappiness`** *(uint64, OPTIONAL)* - sets swappiness parameter of vmscan (See sysctl's vm.swappiness) The values are from 0 to 100. Higher means more swappy.
- **`disableOOMKiller`** *(bool, OPTIONAL)* - enables or disables the OOM killer. If enabled (`false`), tasks that attempt to consume more memory than they are allowed are immediately killed by the OOM killer. The OOM killer is enabled by default in every cgroup using the `memory` subsystem. To disable it, specify a value of `true`.
- **`useHierarchy`** *(bool, OPTIONAL)* - enables or disables hierarchical memory accounting. If enabled (`true`), child cgroups will share the memory limits of this cgroup.
- **`checkBeforeUpdate`** *(bool, OPTIONAL)* - enables container memory usage check before setting a new limit. If enabled (`true`), runtime MAY check if a new memory limit is lower than the current usage, and MUST reject the new limit. Practically, when cgroup v1 is used, the kernel rejects the limit lower than the current usage, and when cgroup v2 is used, an OOM killer is invoked. This setting can be used on cgroup v2 to mimic the cgroup v1 behavior.

Example

```
"memory": {
    "limit": 536870912,
    "reservation": 536870912,
    "swap": 536870912,
    "kernel": -1,
    "kernelTCP": -1,
    "swappiness": 0,
    "disableOOMKiller": false
}
```

CPU

`cpu` (object, OPTIONAL) represents the cgroup subsystems `cpu` and `cpusets`. For more information, see the kernel cgroups documentation about cpusets.

The following parameters can be specified to set up the controller:

- **`shares`** *(uint64, OPTIONAL)* - specifies a relative share of CPU time available to the tasks in a cgroup
- **`quota`** *(int64, OPTIONAL)* - specifies the total amount of time in microseconds for which all tasks in a cgroup can run during one period (as defined by **`period`** below) If specified with any (valid) positive value, it MUST be no smaller than **`burst`** (runtimes MAY generate an error).
- **`burst`** *(uint64, OPTIONAL)* - specifies the maximum amount of accumulated time in microseconds for which all tasks in a cgroup can run additionally for burst during one period (as defined by **`period`** below) If specified, this value MUST be no larger than any positive **`quota`** (runtimes MAY generate an error).
- **`period`** *(uint64, OPTIONAL)* - specifies a period of time in microseconds for how regularly a cgroup's access to CPU resources should be reallocated (CFS scheduler only)
- **`realtimeRuntime`** *(int64, OPTIONAL)* - specifies a period of time in microseconds for the longest continuous period in which the tasks in a cgroup have access to CPU resources
- **`realtimePeriod`** *(uint64, OPTIONAL)* - same as **`period`** but applies to realtime scheduler only
- **`cpus`** *(string, OPTIONAL)* - list of CPUs the container will run in
- **`mems`** *(string, OPTIONAL)* - list of Memory Nodes the container will run in
- **`idle`** *(int64, OPTIONAL)* - cgroups are configured with minimum weight, 0: default behavior, 1: SCHED_IDLE.

Example

```
"cpu": {
    "shares": 1024,
    "quota": 1000000,
    "burst": 1000000,
    "period": 500000,
    "realtimeRuntime": 950000,
    "realtimePeriod": 1000000,
    "cpus": "2-3",
    "mems": "0-7",
    "idle": 0
}
```

Block IO

`blockIO` (object, OPTIONAL) represents the cgroup subsystem **`blkio`** which implements the block IO controller. For more information, see the kernel cgroups documentation about blkio of cgroup v1 or io of cgroup v2, .

Note that I/O throttling settings in cgroup v1 apply only to Direct I/O due to kernel implementation constraints, while this limitation does not exist in cgroup v2.

The following parameters can be specified to set up the controller:

- **`weight`** *(uint16, OPTIONAL)* - specifies per-cgroup weight. This is default weight of the group on all devices until and unless overridden by per-device rules.
- **`leafWeight`** *(uint16, OPTIONAL)* - equivalents of **`weight`** for the purpose of deciding how much weight tasks in the given cgroup has while competing with the cgroup's child cgroups.
- **`weightDevice`** *(array of objects, OPTIONAL)* - an array of per-device bandwidth weights. Each entry has the following structure:
  - **`major, minor`** *(int64, REQUIRED)* - major, minor numbers for device. For more information, see the mknod(1) man page.
  - **`weight`** *(uint16, OPTIONAL)* - bandwidth weight for the device.
  - **`leafWeight`** *(uint16, OPTIONAL)* - bandwidth weight for the device while competing with the cgroup's child cgroups, CFQ scheduler only

  You MUST specify at least one of **`weight`** or **`leafWeight`** in a given entry, and MAY specify both.
- **`throttleReadBpsDevice`**, **`throttleWriteBpsDevice`** *(array of objects, OPTIONAL)* - an array of per-device bandwidth rate limits. Each entry has the following structure:

  - **major, minor** *(int64, REQUIRED)* - major, minor numbers for device. For more information, see the mknod(1) man page.
  - **rate** *(uint64, REQUIRED)* - bandwidth rate limit in bytes per second for the device

- **throttleReadIOPSDevice**, **throttleWriteIOPSDevice** *(array of objects, OPTIONAL)* - an array of per-device IO rate limits. Each entry has the following structure:

  - **major, minor** *(int64, REQUIRED)* - major, minor numbers for device. For more information, see the mknod(1) man page.
  - **rate** *(uint64, REQUIRED)* - IO rate limit for the device

Example

```
"blockIO": {
    "weight": 10,
    "leafWeight": 10,
    "weightDevice": [
        {
            "major": 8,
            "minor": 0,
            "weight": 500,
            "leafWeight": 300
        },
        {
            "major": 8,
            "minor": 16,
            "weight": 500
        }
    ],
    "throttleReadBpsDevice": [
        {
            "major": 8,
            "minor": 0,
            "rate": 600
        }
    ],
    "throttleWriteIOPSDevice": [
        {
            "major": 8,
            "minor": 16,
            "rate": 300
        }
    ]
}
```

Huge page limits

`hugepageLimits` (array of objects, OPTIONAL) represents the `hugetlb` controller which allows to limit the HugeTLB reservations (if supported) or usage (page fault). By default if supported by the kernel, `hugepageLimits` defines the hugepage sizes and limits for HugeTLB controller reservation accounting, which allows to limit the HugeTLB reservations per control group and enforces the controller limit at reservation time and at the fault of HugeTLB memory for which no reservation exists. Otherwise if not supported by the kernel, this should fallback to the page fault accounting, which allows users to limit the HugeTLB usage (page fault) per control group and enforces the limit during page fault.

Note that reservation limits are superior to page fault limits, since reservation limits are enforced at reservation time (on mmap or shget), and never causes the application to get SIGBUS signal if the memory was reserved before hand. This allows for easier fallback to alternatives such as non-HugeTLB memory for example. In the case of page fault accounting, it's very hard to avoid processes getting SIGBUS since the sysadmin needs precisely know the HugeTLB usage of all the tasks in the system and make sure there is enough pages to satisfy all requests. Avoiding tasks getting SIGBUS on overcommited systems is practically impossible with page fault accounting.

For more information, see the kernel cgroups documentation about HugeTLB.

Each entry has the following structure:

- `pageSize` *(string, REQUIRED)* - hugepage size. The value has the format `<size><unit-prefix>B` (64KB, 2MB, 1GB), and must match the `<hugepagesize>` of the corresponding control file found in `/sys/fs/cgroup/hugetlb/hugetlb.<hugepagesize>.rsvd.limit_in_bytes` (if hugetlb_cgroup reservation is supported) or `/sys/fs/cgroup/hugetlb/hugetlb.<hugepagesize>.li` (if not supported). Values of `<unit-prefix>` are intended to be parsed using base 1024 ("1KB" = 1024, "1MB" = 1048576, etc).
- `limit` *(uint64, REQUIRED)* - limit in bytes of *hugepagesize* HugeTLB reservations (if supported) or usage.

Example

```
"hugepageLimits": [
    {
        "pageSize": "2MB",
        "limit": 209715200
    },
    {
        "pageSize": "64KB",
```

```
        "limit": 1000000
    }
]
```

Network

network (object, OPTIONAL) represents the cgroup subsystems **net_cls** and **net_prio**. For more information, see the kernel cgroups documentations about net_cls cgroup and net_prio cgroup.

The following parameters can be specified to set up the controller:

- **classID** *(uint32, OPTIONAL)* - is the network class identifier the cgroup's network packets will be tagged with
- **priorities** *(array of objects, OPTIONAL)* - specifies a list of objects of the priorities assigned to traffic originating from processes in the group and egressing the system on various interfaces. The following parameters can be specified per-priority:
  - **name** *(string, REQUIRED)* - interface name in runtime network namespace
  - **priority** *(uint32, REQUIRED)* - priority applied to the interface

Example

```
"network": {
    "classID": 1048577,
    "priorities": [
        {
            "name": "eth0",
            "priority": 500
        },
        {
            "name": "eth1",
            "priority": 1000
        }
    ]
}
```

PIDs

pids (object, OPTIONAL) represents the cgroup subsystem **pids**. For more information, see the kernel cgroups documentation about pids.

The following parameters can be specified to set up the controller:

- **`limit`** *(int64, REQUIRED)* - specifies the maximum number of tasks in the cgroup

Example

```
"pids": {
    "limit": 32771
}
```

RDMA

`rdma` (object, OPTIONAL) represents the cgroup subsystem **`rdma`**. For more information, see the kernel cgroups documentation about rdma.

The name of the device to limit is the entry key. Entry values are objects with the following properties:

- **`hcaHandles`** *(uint32, OPTIONAL)* - specifies the maximum number of hca_handles in the cgroup
- **`hcaObjects`** *(uint32, OPTIONAL)* - specifies the maximum number of hca_objects in the cgroup

You MUST specify at least one of the **`hcaHandles`** or **`hcaObjects`** in a given entry, and MAY specify both.

Example

```
"rdma": {
    "mlx5_1": {
        "hcaHandles": 3,
        "hcaObjects": 10000
    },
    "mlx4_0": {
        "hcaObjects": 1000
    },
    "rxe3": {
        "hcaObjects": 10000
    }
}
```

Unified

`unified` (object, OPTIONAL) allows cgroup v2 parameters to be to be set and modified for the container.

Each key in the map refers to a file in the cgroup unified hierarchy.

The OCI runtime MUST ensure that the needed cgroup controllers are enabled for the cgroup.

Configuration unknown to the runtime MUST still be written to the relevant file.

The runtime MUST generate an error when the configuration refers to a cgroup controller that is not present or that cannot be enabled.

Example

```
"unified": {
    "io.max": "259:0 rbps=2097152 wiops=120\n253:0 rbps=2097152 wiops=120",
    "hugetlb.1GB.max": "1073741824"
}
```

If a controller is enabled on the cgroup v2 hierarchy but the configuration is provided for the cgroup v1 equivalent controller, the runtime MAY attempt a conversion.

If the conversion is not possible the runtime MUST generate an error.

IntelRdt

`intelRdt` (object, OPTIONAL) represents the Intel Resource Director Technology. If `intelRdt` is set, the runtime MUST write the container process ID to the `tasks` file in a proper sub-directory in a mounted `resctrl` pseudo-filesystem. That sub-directory name is specified by `closID` parameter. If no mounted `resctrl` pseudo-filesystem is available in the runtime mount namespace, the runtime MUST generate an error.

If `intelRdt` is not set, the runtime MUST NOT manipulate any `resctrl` pseudo-filesystems.

The following parameters can be specified for the container:

- `closID` *(string, OPTIONAL)* - specifies the identity for RDT Class of Service (CLOS).

- `l3CacheSchema` *(string, OPTIONAL)* - specifies the schema for L3 cache id and capacity bitmask (CBM). The value SHOULD start with `L3:` and SHOULD NOT contain newlines.
- `memBwSchema` *(string, OPTIONAL)* - specifies the schema of memory bandwidth per L3 cache id. The value MUST start with `MB:` and MUST NOT contain newlines.

The following rules on parameters MUST be applied:

- If both `l3CacheSchema` and `memBwSchema` are set, runtimes MUST write the combined value to the `schemata` file in that sub-directory discussed in `closID`.
- If `l3CacheSchema` contains a line beginning with `MB:`, the value written to `schemata` file MUST be the non-`MB:` line(s) from `l3CacheSchema` and the line from `memBWSchema`.
- If either `l3CacheSchema` or `memBwSchema` is set, runtimes MUST write the value to the `schemata` file in the that sub-directory discussed in `closID`.
- If neither `l3CacheSchema` nor `memBwSchema` is set, runtimes MUST NOT write to `schemata` files in any `resctrl` pseudo-filesystems.
- If `closID` is not set, runtimes MUST use the container ID from `start` and create the `<container-id>` directory.
- If `closID` is set, `l3CacheSchema` and/or `memBwSchema` is set
  - if `closID` directory in a mounted `resctrl` pseudo-filesystem doesn't exist, the runtimes MUST create it.
  - if `closID` directory in a mounted `resctrl` pseudo-filesystem exists, runtimes MUST compare `l3CacheSchema` and/or `memBwSchema` value with `schemata` file, and generate an error if doesn't match.
- If `closID` is set, and neither of `l3CacheSchema` and `memBwSchema` are set, runtime MUST check if corresponding pre-configured directory `closID` is present in mounted `resctrl`. If such pre-configured directory `closID` exists, runtime MUST assign container to this `closID` and generate an error if directory does not exist.
- `enableCMT` *(boolean, OPTIONAL)* - specifies if Intel RDT CMT should be enabled:
  - CMT (Cache Monitoring Technology) supports monitoring of the last-level cache (LLC) occupancy for the container.
- `enableMBM` *(boolean, OPTIONAL)* - specifies if Intel RDT MBM should be enabled:

– MBM (Memory Bandwidth Monitoring) supports monitoring of total and local memory bandwidth for the container.

Example

Consider a two-socket machine with two L3 caches where the default CBM is 0x7ff and the max CBM length is 11 bits, and minimum memory bandwidth of 10% with a memory bandwidth granularity of 10%.

Tasks inside the container only have access to the "upper" 7/11 of L3 cache on socket 0 and the "lower" 5/11 L3 cache on socket 1, and may use a maximum memory bandwidth of 20% on socket 0 and 70% on socket 1.

```
"linux": {
    "intelRdt": {
        "closID": "guaranteed_group",
        "l3CacheSchema": "L3:0=7f0;1=1f",
        "memBwSchema": "MB:0=20;1=70"
    }
}
```

Sysctl

sysctl (object, OPTIONAL) allows kernel parameters to be modified at runtime for the container. For more information, see the sysctl(8) man page.

Example

```
"sysctl": {
    "net.ipv4.ip_forward": "1",
    "net.core.somaxconn": "256"
}
```

Seccomp

Seccomp provides application sandboxing mechanism in the Linux kernel. Seccomp configuration allows one to configure actions to take for matched syscalls and furthermore also allows matching on values passed as arguments to syscalls. For more information about Seccomp, see Seccomp kernel documentation. The actions, architectures, and operators are strings that match the definitions in seccomp.h from libseccomp and are translated to corresponding values.

seccomp (object, OPTIONAL)

The following parameters can be specified to set up seccomp:

- **`defaultAction`** *(string, REQUIRED)* - the default action for seccomp. Allowed values are the same as `syscalls[].action`.
- **`defaultErrnoRet`** *(uint, OPTIONAL)* - the errno return code to use. Some actions like `SCMP_ACT_ERRNO` and `SCMP_ACT_TRACE` allow to specify the errno code to return. When the action doesn't support an errno, the runtime MUST print and error and fail. If not specified then its default value is `EPERM`.
- **`architectures`** *(array of strings, OPTIONAL)* - the architecture used for system calls. A valid list of constants as of libseccomp v2.5.0 is shown below.
  - `SCMP_ARCH_X86`
  - `SCMP_ARCH_X86_64`
  - `SCMP_ARCH_X32`
  - `SCMP_ARCH_ARM`
  - `SCMP_ARCH_AARCH64`
  - `SCMP_ARCH_MIPS`
  - `SCMP_ARCH_MIPS64`
  - `SCMP_ARCH_MIPS64N32`
  - `SCMP_ARCH_MIPSEL`
  - `SCMP_ARCH_MIPSEL64`
  - `SCMP_ARCH_MIPSEL64N32`
  - `SCMP_ARCH_PPC`
  - `SCMP_ARCH_PPC64`
  - `SCMP_ARCH_PPC64LE`
  - `SCMP_ARCH_S390`
  - `SCMP_ARCH_S390X`
  - `SCMP_ARCH_PARISC`
  - `SCMP_ARCH_PARISC64`
  - `SCMP_ARCH_RISCV64`
- **`flags`** *(array of strings, OPTIONAL)* - list of flags to use with seccomp(2). A valid list of constants is shown below.
  - `SECCOMP_FILTER_FLAG_TSYNC`
  - `SECCOMP_FILTER_FLAG_LOG`
  - `SECCOMP_FILTER_FLAG_SPEC_ALLOW`
  - `SECCOMP_FILTER_FLAG_WAIT_KILLABLE_RECV`
- **`listenerPath`** *(string, OPTIONAL)* - specifies the path of UNIX domain socket over which the runtime will send the container process state data structure when the `SCMP_ACT_NOTIFY` action is used. This socket MUST use `AF_UNIX` domain and `SOCK_STREAM` type. The runtime MUST send exactly one container process state per connection. The connection MUST NOT be reused and it MUST be closed after sending a seccomp state. If

sending to this socket fails, the runtime MUST generate an error. If the `SCMP_ACT_NOTIFY` action is not used this value is ignored.

The runtime sends the following file descriptors using `SCM_RIGHTS` and set their names in the `fds` array of the container process state:

  - **`seccompFd`** (string, REQUIRED) is the seccomp file descriptor returned by the seccomp syscall.

- **`listenerMetadata`** *(string, OPTIONAL)* - specifies an opaque data to pass to the seccomp agent. This string will be sent as the `metadata` field in the container process state. This field MUST NOT be set if `listenerPath` is not set.

- **`syscalls`** *(array of objects, OPTIONAL)* - match a syscall in seccomp. While this property is OPTIONAL, some values of `defaultAction` are not useful without `syscalls` entries. For example, if `defaultAction` is `SCMP_ACT_KILL` and `syscalls` is empty or unset, the kernel will kill the container process on its first syscall. Each entry has the following structure:

  - **`names`** *(array of strings, REQUIRED)* - the names of the syscalls. `names` MUST contain at least one entry.
  - **`action`** *(string, REQUIRED)* - the action for seccomp rules. A valid list of constants as of libseccomp v2.5.0 is shown below.
    * `SCMP_ACT_KILL`
    * `SCMP_ACT_KILL_PROCESS`
    * `SCMP_ACT_KILL_THREAD`
    * `SCMP_ACT_TRAP`
    * `SCMP_ACT_ERRNO`
    * `SCMP_ACT_TRACE`
    * `SCMP_ACT_ALLOW`
    * `SCMP_ACT_LOG`
    * `SCMP_ACT_NOTIFY`
  - **`errnoRet`** *(uint, OPTIONAL)* - the errno return code to use. Some actions like `SCMP_ACT_ERRNO` and `SCMP_ACT_TRACE` allow to specify the errno code to return. When the action doesn't support an errno, the runtime MUST print and error and fail. If not specified its default value is `EPERM`.
  - **`args`** *(array of objects, OPTIONAL)* - the specific syscall in seccomp. Each entry has the following structure:
    * **`index`** *(uint, REQUIRED)* - the index for syscall arguments in seccomp.
    * **`value`** *(uint64, REQUIRED)* - the value for syscall arguments in seccomp.

* `valueTwo` *(uint64, OPTIONAL)* - the value for syscall arguments in seccomp.
* `op` *(string, REQUIRED)* - the operator for syscall arguments in seccomp. A valid list of constants as of libseccomp v2.3.2 is shown below.
  - SCMP_CMP_NE
  - SCMP_CMP_LT
  - SCMP_CMP_LE
  - SCMP_CMP_EQ
  - SCMP_CMP_GE
  - SCMP_CMP_GT
  - SCMP_CMP_MASKED_EQ

Example

```
"seccomp": {
    "defaultAction": "SCMP_ACT_ALLOW",
    "architectures": [
        "SCMP_ARCH_X86",
        "SCMP_ARCH_X32"
    ],
    "syscalls": [
        {
            "names": [
                "getcwd",
                "chmod"
            ],
            "action": "SCMP_ACT_ERRNO"
        }
    ]
}
```

The Container Process State

The container process state is a data structure passed via a UNIX socket. The container runtime MUST send the container process state over the UNIX socket as regular payload serialized in JSON and file descriptors MUST be sent using `SCM_RIGHTS`. The container runtime MAY use several `sendmsg(2)` calls to send the aforementioned data. If more than one `sendmsg(2)` is used, the file descriptors MUST be sent only in the first call.

The container process state includes the following properties:

- **ociVersion** (string, REQUIRED) is version of the Open Container Initiative Runtime Specification with which the container process state complies.
- **fds** (array, OPTIONAL) is a string array containing the names of the file descriptors passed. The index of the name in this array corresponds to index of the file descriptors in the SCM_RIGHTS array.
- **pid** (int, REQUIRED) is the container process ID, as seen by the runtime.
- **metadata** (string, OPTIONAL) opaque metadata.
- **state** (state, REQUIRED) is the state of the container.

Example sending a single seccompFd file descriptor in the SCM_RIGHTS array:

```
{
    "ociVersion": "1.0.2",
    "fds": [
        "seccompFd"
    ],
    "pid": 4422,
    "metadata": "MKNOD=/dev/null,/dev/net/tun;BPF_MAP_TYPES=hash,array",
    "state": {
        "ociVersion": "1.0.2",
        "id": "oci-container1",
        "status": "creating",
        "pid": 4422,
        "bundle": "/containers/redis",
        "annotations": {
            "myKey": "myValue"
        }
    }
}
```

Rootfs Mount Propagation

rootfsPropagation (string, OPTIONAL) sets the rootfs's mount propagation. Its value is either **shared**, **slave**, **private** or **unbindable**. It's worth noting that a peer group is defined as a group of VFS mounts that propagate events to each other. A nested container is defined as a container launched inside an existing container.

- **shared**: the rootfs mount belongs to a new peer group. This means that further mounts (e.g. nested containers) will also belong to that peer group and will propagate events to the rootfs. Note this does not mean that it's shared with the host.

- **slave**: the rootfs mount receives propagation events from the host (e.g. if something is mounted on the host it will also appear in the container) but not the other way around.
- **private**: the rootfs mount doesn't receive mount propagation events from the host and further mounts in nested containers will be isolated from the host and from the rootfs (even if the nested container `rootfsPropagation` option is shared).
- **unbindable**: the rootfs mount is a private mount that cannot be bind-mounted.

The Shared Subtrees article in the kernel documentation has more information about mount propagation.

Example

```
"rootfsPropagation": "slave",
```

Masked Paths

maskedPaths (array of strings, OPTIONAL) will mask over the provided paths inside the container so that they cannot be read. The values MUST be absolute paths in the container namespace.

Example

```
"maskedPaths": [
    "/proc/kcore"
]
```

Readonly Paths

readonlyPaths (array of strings, OPTIONAL) will set the provided paths as readonly inside the container. The values MUST be absolute paths in the container namespace.

Example

```
"readonlyPaths": [
    "/proc/sys"
]
```

Mount Label

`mountLabel` (string, OPTIONAL) will set the Selinux context for the mounts in the container.

Example

```
"mountLabel": "system_u:object_r:svirt_sandbox_file_t:s0:c715,c811"
```

Personality

`personality` (object, OPTIONAL) sets the Linux execution personality. For more information see the personality syscall documentation. As most of the options are obsolete and rarely used, and some reduce security, the currently supported set is a small subset of the available options.

- **`domain`** *(string, REQUIRED)* - the execution domain. The valid list of constants is shown below. `LINUX32` will set the `uname` system call to show a 32 bit CPU type, such as `i686`.
  - `LINUX`
  - `LINUX32`
- **`flags`** *(array of strings, OPTIONAL)* - the additional flags to apply. Currently no flag values are supported.

Solaris Application Container Configuration

Solaris application containers can be configured using the following properties, all of the below properties have mappings to properties specified under zonecfg(1M) man page, except milestone.

milestone

The SMF(Service Management Facility) FMRI which should go to "online" state before we start the desired process within the container.

`milestone` *(string, OPTIONAL)*

Example

```
"milestone": "svc:/milestone/container:default"
```

limitpriv

The maximum set of privileges any process in this container can obtain. The property should consist of a comma-separated privilege set specification as described in priv_str_to_set(3C) man page for the respective release of Solaris.

limitpriv *(string, OPTIONAL)*

Example

```
"limitpriv": "default"
```

maxShmMemory

The maximum amount of shared memory allowed for this application container. A scale (K, M, G, T) can be applied to the value for each of these numbers (for example, 1M is one megabyte). Mapped to **max-shm-memory** in zonecfg(1M) man page.

maxShmMemory *(string, OPTIONAL)*

Example

```
"maxShmMemory": "512m"
```

cappedCPU

Sets a limit on the amount of CPU time that can be used by a container. The unit used translates to the percentage of a single CPU that can be used by all user threads in a container, expressed as a fraction (for example, .75) or a mixed number (whole number and fraction, for example, 1.25). An ncpu value of 1 means 100% of a CPU, a value of 1.25 means 125%, .75 mean 75%, and so forth. When projects within a capped container have their own caps, the minimum value takes precedence. cappedCPU is mapped to **capped-cpu** in zonecfg(1M) man page.

- **ncpus** *(string, OPTIONAL)*

Example

```
"cappedCPU": {
    "ncpus": "8"
}
```

cappedMemory

The physical and swap caps on the memory that can be used by this application container. A scale (K, M, G, T) can be applied to the value for each of these numbers (for example, 1M is one megabyte). cappedMemory is mapped to `capped-memory` in zonecfg(1M) man page.

- **`physical`** *(string, OPTIONAL)*
- **`swap`** *(string, OPTIONAL)*

Example

```
"cappedMemory": {
    "physical": "512m",
    "swap": "512m"
}
```

Network

Automatic Network (anet)

anet is specified as an array that is used to set up networking for Solaris application containers. The anet resource represents the automatic creation of a network resource for an application container. The zones administration daemon, zoneadmd, is the primary process for managing the container's virtual platform. One of the daemon's responsibilities is creation and teardown of the networks for the container. For more information on the daemon see the zoneadmd(1M) man page. When such a container is started, a temporary VNIC(Virtual NIC) is automatically created for the container. The VNIC is deleted when the container is torn down. The following properties can be used to set up automatic networks. For additional information on properties, check the zonecfg(1M) man page for the respective release of Solaris.

- **`linkname`** *(string, OPTIONAL)* Specify a name for the automatically created VNIC datalink.
- **`lowerLink`** *(string, OPTIONAL)* Specify the link over which the VNIC will be created. Mapped to `lower-link` in the zonecfg(1M) man page.
- **`allowedAddress`** *(string, OPTIONAL)* The set of IP addresses that the container can use might be constrained by specifying the `allowedAddress` property. If `allowedAddress` has not been specified, then they can use any IP address on the associated physical interface for the network resource. Otherwise, when `allowedAddress` is specified, the container cannot use IP addresses that are not in the `allowedAddress` list for the physical address. Mapped to `allowed-address` in the zonecfg(1M) man page.

- `configureAllowedAddress` *(string, OPTIONAL)* If `configureAllowedAddress` is set to true, the addresses specified by `allowedAddress` are automatically configured on the interface each time the container starts. When it is set to false, the `allowedAddress` will not be configured on container start. Mapped to `configure-allowed-address` in the zonecfg(1M) man page.
- `defrouter` *(string, OPTIONAL)* The value for the OPTIONAL default router.
- `macAddress` *(string, OPTIONAL)* Set the VNIC's MAC addresses based on the specified value or keyword. If not a keyword, it is interpreted as a unicast MAC address. For a list of the supported keywords please refer to the zonecfg(1M) man page of the respective Solaris release. Mapped to `mac-address` in the zonecfg(1M) man page.
- `linkProtection` *(string, OPTIONAL)* Enables one or more types of link protection using comma-separated values. See the protection property in dladm(8) for supported values in respective release of Solaris. Mapped to `link-protection` in the zonecfg(1M) man page.

Example

```
"anet": [
    {
        "allowedAddress": "172.17.0.2/16",
        "configureAllowedAddress": "true",
        "defrouter": "172.17.0.1/16",
        "linkProtection": "mac-nospoof, ip-nospoof",
        "linkname": "net0",
        "lowerLink": "net2",
        "macAddress": "02:42:f8:52:c7:16"
    }
]
```

Features Structure

A runtime MAY provide a JSON structure about its implemented features to runtime callers. This JSON structure is called "Features structure".

The Features structure is irrelevant to the actual availability of the features in the host operating system. Hence, the content of the Features structure SHOULD be determined on the compilation time of the runtime, not on the execution time.

All properties in the Features structure except `ociVersionMin` and `ociVersionMax` MAY either be absent or have the `null` value. The `null` value

MUST NOT be confused with an empty value such as `0`, `false`, `""`, `[]`, and `{}`.

Specification version

- **`ociVersionMin`** (string, REQUIRED) The minimum recognized version of the Open Container Initiative Runtime Specification. The runtime MUST accept this value as the `ociVersion` property of `config.json`.
- **`ociVersionMax`** (string, REQUIRED) The maximum recognized version of the Open Container Initiative Runtime Specification. The runtime MUST accept this value as the `ociVersion` property of `config.json`. The value MUST NOT be less than the value of the `ociVersionMin` property. The Features structure MUST NOT contain properties that are not defined in this version of the Open Container Initiative Runtime Specification.

Example

```
{
  "ociVersionMin": "1.0.0",
  "ociVersionMax": "1.1.0"
}
```

Hooks

- **`hooks`** (array of strings, OPTIONAL) The recognized names of the hooks. The runtime MUST support the elements in this array as the `hooks` property of `config.json`.

Example

```
"hooks": [
  "prestart",
  "createRuntime",
  "createContainer",
  "startContainer",
  "poststart",
  "poststop"
]
```

Mount Options

- **`mountOptions`** (array of strings, OPTIONAL) The recognized names of the mount options, including options that might not be supported by the host operating system. The runtime MUST recognize the elements in this array as the `options` of `mounts` objects in `config.json`.
  - Linux: this array SHOULD NOT contain filesystem-specific mount options that are passed to the mount(2) syscall as `const void *data`.

Example

```
"mountOptions": [
  "acl",
  "async",
  "atime",
  "bind",
  "defaults",
  "dev",
  "diratime",
  "dirsync",
  "exec",
  "iversion",
  "lazytime",
  "loud",
  "mand",
  "noacl",
  "noatime",
  "nodev",
  "nodiratime",
  "noexec",
  "noiversion",
  "nolazytime",
  "nomand",
  "norelatime",
  "nostrictatime",
  "nosuid",
  "nosymfollow",
  "private",
  "ratime",
  "rbind",
  "rdev",
  "rdiratime",
  "relatime",
  "remount",
```

```
    "rexec",
    "rnoatime",
    "rnodev",
    "rnodiratime",
    "rnoexec",
    "rnorelatime",
    "rnostrictatime",
    "rnosuid",
    "rnosymfollow",
    "ro",
    "rprivate",
    "rrelatime",
    "rro",
    "rrw",
    "rshared",
    "rslave",
    "rstrictatime",
    "rsuid",
    "rsymfollow",
    "runbindable",
    "rw",
    "shared",
    "silent",
    "slave",
    "strictatime",
    "suid",
    "symfollow",
    "sync",
    "tmpcopyup",
    "unbindable"
]
```

Platform-specific features

- **`linux`** (object, OPTIONAL) Linux-specific features. This MAY be set if the runtime supports `linux` platform.

Annotations

`annotations` (object, OPTIONAL) contains arbitrary metadata of the runtime. This information MAY be structured or unstructured. Annotations MUST be a key-value map that follows the same convention as the Key and Values of the `annotations` property of `config.json`. However, annotations do not need to contain the possible values of the `annotations` property of `config.json`.

The current version of the spec do not provide a way to enumerate the possible values of the `annotations` property of `config.json`.

Example

```
"annotations": {
  "org.opencontainers.runc.checkpoint.enabled": "true",
  "org.opencontainers.runc.version": "1.1.0"
}
```

Unsafe annotations in `config.json`

`potentiallyUnsafeConfigAnnotations` (array of strings, OPTIONAL) contains values of `annotations` property of `config.json` that may potentially change the behavior of the runtime.

A value that ends with "." is interpreted as a prefix of annotations.

Example

```
"potentiallyUnsafeConfigAnnotations": [
  "com.example.foo.bar",
  "org.systemd.property."
]
```

The example above matches `com.example.foo.bar`, `org.systemd.property.ExecStartPre`, etc. The example does not match `com.example.foo.bar.baz`.

Example

Here is a full example for reference.

```
{
  "ociVersionMin": "1.0.0",
  "ociVersionMax": "1.1.0-rc.2",
  "hooks": [
    "prestart",
    "createRuntime",
    "createContainer",
    "startContainer",
    "poststart",
    "poststop"
```

```
],
"mountOptions": [
  "async",
  "atime",
  "bind",
  "defaults",
  "dev",
  "diratime",
  "dirsync",
  "exec",
  "iversion",
  "lazytime",
  "loud",
  "mand",
  "noatime",
  "nodev",
  "nodiratime",
  "noexec",
  "noiversion",
  "nolazytime",
  "nomand",
  "norelatime",
  "nostrictatime",
  "nosuid",
  "nosymfollow",
  "private",
  "ratime",
  "rbind",
  "rdev",
  "rdiratime",
  "relatime",
  "remount",
  "rexec",
  "rnoatime",
  "rnodev",
  "rnodiratime",
  "rnoexec",
  "rnorelatime",
  "rnostrictatime",
  "rnosuid",
  "rnosymfollow",
  "ro",
  "rprivate",
  "rrelatime",
  "rro",
  "rrw",
```

```
      "rshared",
      "rslave",
      "rstrictatime",
      "rsuid",
      "rsymfollow",
      "runbindable",
      "rw",
      "shared",
      "silent",
      "slave",
      "strictatime",
      "suid",
      "symfollow",
      "sync",
      "tmpcopyup",
      "unbindable"
    ],
    "linux": {
      "namespaces": [
        "cgroup",
        "ipc",
        "mount",
        "network",
        "pid",
        "user",
        "uts"
      ],
      "capabilities": [
        "CAP_CHOWN",
        "CAP_DAC_OVERRIDE",
        "CAP_DAC_READ_SEARCH",
        "CAP_FOWNER",
        "CAP_FSETID",
        "CAP_KILL",
        "CAP_SETGID",
        "CAP_SETUID",
        "CAP_SETPCAP",
        "CAP_LINUX_IMMUTABLE",
        "CAP_NET_BIND_SERVICE",
        "CAP_NET_BROADCAST",
        "CAP_NET_ADMIN",
        "CAP_NET_RAW",
        "CAP_IPC_LOCK",
        "CAP_IPC_OWNER",
        "CAP_SYS_MODULE",
        "CAP_SYS_RAWIO",
```

```
    "CAP_SYS_CHROOT",
    "CAP_SYS_PTRACE",
    "CAP_SYS_PACCT",
    "CAP_SYS_ADMIN",
    "CAP_SYS_BOOT",
    "CAP_SYS_NICE",
    "CAP_SYS_RESOURCE",
    "CAP_SYS_TIME",
    "CAP_SYS_TTY_CONFIG",
    "CAP_MKNOD",
    "CAP_LEASE",
    "CAP_AUDIT_WRITE",
    "CAP_AUDIT_CONTROL",
    "CAP_SETFCAP",
    "CAP_MAC_OVERRIDE",
    "CAP_MAC_ADMIN",
    "CAP_SYSLOG",
    "CAP_WAKE_ALARM",
    "CAP_BLOCK_SUSPEND",
    "CAP_AUDIT_READ",
    "CAP_PERFMON",
    "CAP_BPF",
    "CAP_CHECKPOINT_RESTORE"
  ],
  "cgroup": {
    "v1": true,
    "v2": true,
    "systemd": true,
    "systemdUser": true,
    "rdma": true
  },
  "seccomp": {
    "enabled": true,
    "actions": [
      "SCMP_ACT_ALLOW",
      "SCMP_ACT_ERRNO",
      "SCMP_ACT_KILL",
      "SCMP_ACT_KILL_PROCESS",
      "SCMP_ACT_KILL_THREAD",
      "SCMP_ACT_LOG",
      "SCMP_ACT_NOTIFY",
      "SCMP_ACT_TRACE",
      "SCMP_ACT_TRAP"
    ],
    "operators": [
      "SCMP_CMP_EQ",
```

```
      "SCMP_CMP_GE",
      "SCMP_CMP_GT",
      "SCMP_CMP_LE",
      "SCMP_CMP_LT",
      "SCMP_CMP_MASKED_EQ",
      "SCMP_CMP_NE"
    ],
    "archs": [
      "SCMP_ARCH_AARCH64",
      "SCMP_ARCH_ARM",
      "SCMP_ARCH_MIPS",
      "SCMP_ARCH_MIPS64",
      "SCMP_ARCH_MIPS64N32",
      "SCMP_ARCH_MIPSEL",
      "SCMP_ARCH_MIPSEL64",
      "SCMP_ARCH_MIPSEL64N32",
      "SCMP_ARCH_PPC",
      "SCMP_ARCH_PPC64",
      "SCMP_ARCH_PPC64LE",
      "SCMP_ARCH_RISCV64",
      "SCMP_ARCH_S390",
      "SCMP_ARCH_S390X",
      "SCMP_ARCH_X32",
      "SCMP_ARCH_X86",
      "SCMP_ARCH_X86_64"
    ],
    "knownFlags": [
      "SECCOMP_FILTER_FLAG_TSYNC",
      "SECCOMP_FILTER_FLAG_SPEC_ALLOW",
      "SECCOMP_FILTER_FLAG_LOG"
    ],
    "supportedFlags": [
      "SECCOMP_FILTER_FLAG_TSYNC",
      "SECCOMP_FILTER_FLAG_SPEC_ALLOW",
      "SECCOMP_FILTER_FLAG_LOG"
    ]
  },
  "apparmor": {
    "enabled": true
  },
  "selinux": {
    "enabled": true
  },
  "intelRdt": {
    "enabled": true
  }
```

```
  },
  "annotations": {
    "io.github.seccomp.libseccomp.version": "2.5.4",
    "org.opencontainers.runc.checkpoint.enabled": "true",
    "org.opencontainers.runc.commit": "v1.1.0-534-g26851168",
    "org.opencontainers.runc.version": "1.1.0+dev"
  }
}
```

Linux Features Structure

This document describes the Linux-specific section of the Features structure.

Namespaces

- **`namespaces`** (array of strings, OPTIONAL) The recognized names of the namespaces, including namespaces that might not be supported by the host operating system. The runtime MUST recognize the elements in this array as the `type` of `linux.namespaces` objects in `config.json`.

Example

```
"namespaces": [
  "cgroup",
  "ipc",
  "mount",
  "network",
  "pid",
  "user",
  "uts"
]
```

Capabilities

- **`capabilities`** (array of strings, OPTIONAL) The recognized names of the capabilities, including capabilities that might not be supported by the host operating system. The runtime MUST recognize the elements in this array in the `process.capabilities` object of `config.json`.

Example

```
"capabilities": [
  "CAP_CHOWN",
  "CAP_DAC_OVERRIDE",
  "CAP_DAC_READ_SEARCH",
  "CAP_FOWNER",
  "CAP_FSETID",
  "CAP_KILL",
  "CAP_SETGID",
  "CAP_SETUID",
  "CAP_SETPCAP",
  "CAP_LINUX_IMMUTABLE",
  "CAP_NET_BIND_SERVICE",
  "CAP_NET_BROADCAST",
  "CAP_NET_ADMIN",
  "CAP_NET_RAW",
  "CAP_IPC_LOCK",
  "CAP_IPC_OWNER",
  "CAP_SYS_MODULE",
  "CAP_SYS_RAWIO",
  "CAP_SYS_CHROOT",
  "CAP_SYS_PTRACE",
  "CAP_SYS_PACCT",
  "CAP_SYS_ADMIN",
  "CAP_SYS_BOOT",
  "CAP_SYS_NICE",
  "CAP_SYS_RESOURCE",
  "CAP_SYS_TIME",
  "CAP_SYS_TTY_CONFIG",
  "CAP_MKNOD",
  "CAP_LEASE",
  "CAP_AUDIT_WRITE",
  "CAP_AUDIT_CONTROL",
  "CAP_SETFCAP",
  "CAP_MAC_OVERRIDE",
  "CAP_MAC_ADMIN",
  "CAP_SYSLOG",
  "CAP_WAKE_ALARM",
  "CAP_BLOCK_SUSPEND",
  "CAP_AUDIT_READ",
  "CAP_PERFMON",
  "CAP_BPF",
  "CAP_CHECKPOINT_RESTORE"
]
```

Cgroup

cgroup (object, OPTIONAL) represents the runtime's implementation status of cgroup managers. Irrelevant to the cgroup version of the host operating system.

- **v1** (bool, OPTIONAL) represents whether the runtime supports cgroup v1.
- **v2** (bool, OPTIONAL) represents whether the runtime supports cgroup v2.
- **systemd** (bool, OPTIONAL) represents whether the runtime supports system-wide systemd cgroup manager.
- **systemdUser** (bool, OPTIONAL) represents whether the runtime supports user-scoped systemd cgroup manager.
- **rdma** (bool, OPTIONAL) represents whether the runtime supports RDMA cgroup controller.

Example

```
"cgroup": {
  "v1": true,
  "v2": true,
  "systemd": true,
  "systemdUser": true,
  "rdma": false
}
```

Seccomp

seccomp (object, OPTIONAL) represents the runtime's implementation status of seccomp. Irrelevant to the kernel version of the host operating system.

- **enabled** (bool, OPTIONAL) represents whether the runtime supports seccomp.
- **actions** (array of strings, OPTIONAL) The recognized names of the seccomp actions. The runtime MUST recognize the elements in this array in the `syscalls[].action` property of the `linux.seccomp` object in `config.json`.
- **operators** (array of strings, OPTIONAL) The recognized names of the seccomp operators. The runtime MUST recognize the elements in this array in the `syscalls[].args[].op` property of the `linux.seccomp` object in `config.json`.
- **archs** (array of strings, OPTIONAL) The recognized names of the seccomp architectures. The runtime MUST recognize the elements in this

array in the `architectures` property of the `linux.seccomp` object in `config.json`.

- **`knownFlags`** (array of strings, OPTIONAL) The recognized names of the seccomp flags. The runtime MUST recognize the elements in this array in the `flags` property of the `linux.seccomp` object in `config.json`.
- **`supportedFlags`** (array of strings, OPTIONAL) The recognized and supported names of the seccomp flags. This list may be a subset of `knownFlags` due to some flags not supported by the current kernel and/or libseccomp. The runtime MUST recognize and support the elements in this array in the `flags` property of the `linux.seccomp` object in `config.json`.

Example

```
"seccomp": {
  "enabled": true,
  "actions": [
    "SCMP_ACT_ALLOW",
    "SCMP_ACT_ERRNO",
    "SCMP_ACT_KILL",
    "SCMP_ACT_LOG",
    "SCMP_ACT_NOTIFY",
    "SCMP_ACT_TRACE",
    "SCMP_ACT_TRAP"
  ],
  "operators": [
    "SCMP_CMP_EQ",
    "SCMP_CMP_GE",
    "SCMP_CMP_GT",
    "SCMP_CMP_LE",
    "SCMP_CMP_LT",
    "SCMP_CMP_MASKED_EQ",
    "SCMP_CMP_NE"
  ],
  "archs": [
    "SCMP_ARCH_AARCH64",
    "SCMP_ARCH_ARM",
    "SCMP_ARCH_MIPS",
    "SCMP_ARCH_MIPS64",
    "SCMP_ARCH_MIPS64N32",
    "SCMP_ARCH_MIPSEL",
    "SCMP_ARCH_MIPSEL64",
    "SCMP_ARCH_MIPSEL64N32",
    "SCMP_ARCH_PPC",
    "SCMP_ARCH_PPC64",
```

```
    "SCMP_ARCH_PPC64LE",
    "SCMP_ARCH_S390",
    "SCMP_ARCH_S390X",
    "SCMP_ARCH_X32",
    "SCMP_ARCH_X86",
    "SCMP_ARCH_X86_64"
  ],
  "knownFlags": [
    "SECCOMP_FILTER_FLAG_LOG"
  ],
  "supportedFlags": [
    "SECCOMP_FILTER_FLAG_LOG"
  ]
}
```

AppArmor

apparmor (object, OPTIONAL) represents the runtime's implementation status of AppArmor. Irrelevant to the availability of AppArmor on the host operating system.

- **enabled** (bool, OPTIONAL) represents whether the runtime supports AppArmor.

Example

```
"apparmor": {
  "enabled": true
}
```

SELinux

selinux (object, OPTIONAL) represents the runtime's implementation status of SELinux. Irrelevant to the availability of SELinux on the host operating system.

- **enabled** (bool, OPTIONAL) represents whether the runtime supports SELinux.

Example

```
"selinux": {
  "enabled": true
}
```

Intel RDT

`intelRdt` (object, OPTIONAL) represents the runtime's implementation status of Intel RDT. Irrelevant to the availability of Intel RDT on the host operating system.

- `enabled` (bool, OPTIONAL) represents whether the runtime supports Intel RDT.

Example

```
"intelRdt": {
  "enabled": true
}
```

MountExtensions

`mountExtensions` (object, OPTIONAL) represents whether the runtime supports certain mount features, irrespective of the availability of the features on the host operating system.

- `idmap` (object, OPTIONAL) represents whether the runtime supports idmap mounts using the `uidMappings` and `gidMappings` properties of the mount.
  - `enabled` (bool, OPTIONAL) represents whether the runtime parses and attempts to use the `uidMappings` and `gidMappings` properties of mounts if provided. Note that it is possible for runtimes to have partial implementations of id-mapped mounts support (such as only allowing mounts which have mappings matching the container's user namespace, or only allowing the id-mapped bind-mounts). In such cases, runtimes MUST still set this value to `true`, to indicate that the runtime recognises the `uidMappings` and `gidMappings` properties.

Example

```
"mountExtensions": {
  "idmap":{
    "enabled": true
  }
}
```

Glossary

Bundle

A directory structure that is written ahead of time, distributed, and used to seed the runtime for creating a container and launching a process within it.

Configuration

The `config.json` file in a bundle which defines the intended container and container process.

Container

An environment for executing processes with configurable isolation and resource limitations. For example, namespaces, resource limits, and mounts are all part of the container environment.

Container namespace

On Linux,the namespaces in which the configured process executes.

Features Structure

A JSON structure that represents the implemented features of the runtime. Irrelevant to the actual availability of the features in the host operating system.

JSON

All configuration JSON MUST be encoded in UTF-8. JSON objects MUST NOT include duplicate names. The order of entries in JSON objects is not significant.

Runtime

An implementation of this specification. It reads the configuration files from a bundle, uses that information to create a container, launches a process inside the container, and performs other lifecycle actions.

Runtime caller

An external program to execute a runtime, directly or indirectly.

Examples of direct callers include containerd, CRI-O, and Podman. Examples of indirect callers include Docker/Moby and Kubernetes.

Runtime callers often execute a runtime via runc-compatible command line interface, however, its interaction interface is currently out of the scope of the Open Container Initiative Runtime Specification.

Runtime namespace

On Linux, the namespaces from which new container namespaces are created and from which some configured resources are accessed.