

Virtualización y Contenedores en Cómputo Científico

Hermilo Cortés González

Escuela de Gobierno y Transformación Pública
Data Pub

27 de mayo de 2024



Centros de Super Cómputo

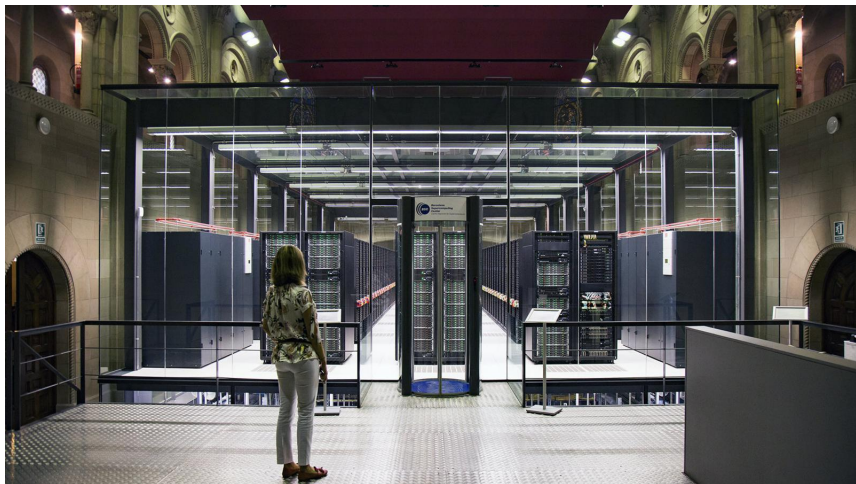
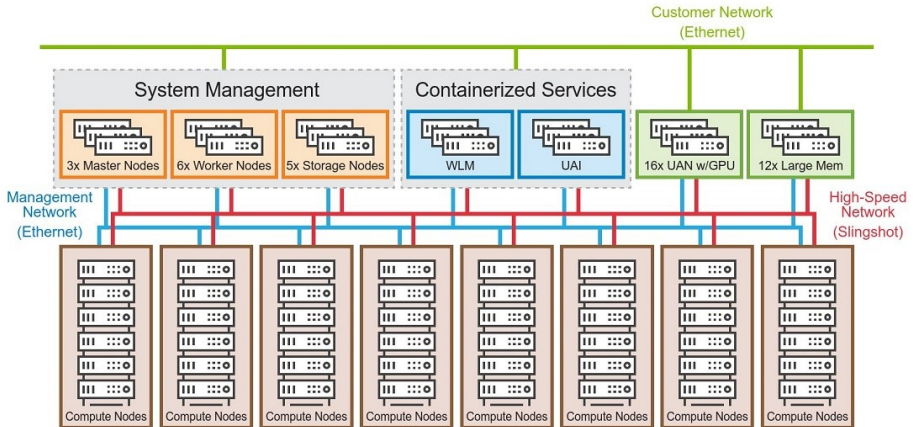


Figura: Barcelona Supercomputing Center

Centros de Super Cómputo : Recurso compartido

- **Usuaría-Clienta.** Relacionadas con un grupo o institución.
- **Administradoras.** Satisfacen distintas necesidades de las clientas:
 - ▶ Versiones de software.
 - ▶ Compiladores.
 - ▶ Módulos.
 - ▶ Sistemas de archivos.
 - ▶ Permisos.

Arquitectura de un Cluster de HPC (High Performance Computing)



HPC Architecture Connection Diagram

Figura: Tomado de <https://www.gigabyte.com/Article/high-performance-computing-cluster>

[//www.gigabyte.com/Article/high-performance-computing-cluster](https://www.gigabyte.com/Article/high-performance-computing-cluster)

¿Cómo se administra un Cluster con miles de nodos?

Herramientas de configuración automatizada



Job Scheduler o Resource Management System(RMS)



Dos visiones

Administradoras

Se aseguran que la clienta-usuario tenga las herramientas y soporte necesario para el uso eficiente de los recursos.



Usuarios

Consume los recursos.



Todo parece estar bien ...



La realidad puede ser otra...



Problema

- Como en todo aquello donde existe el uso de un recurso compartido, las relaciones Sysadmin-Usuaría, Usuaría-Usuaría no son siempre cordiales.
- Hay conflicto : ¿Quién y cuándo tiene acceso al recurso? ¿Hay asignación justa de recursos? ¿Hay usuarios con mayores privilegios?
- **Sysadmin** : Definen **Reglas-Políticas de uso** con el objetivo de satisfacer las necesidades de las usuarias pero también acorde a las necesidades de mantener un recurso usable y confiable.
- **Usuaría** : Estas reglas y políticas se traduce en sistemas y software limitantes e inmóviles (Kurtzer et al., 2017).

Problema

This static nature coupled with distribution-specific software builds meant that service providers would ultimately end up limiting the scope of computational science that their systems could support (Kurtzer et al., 2017)

Propuesta de solución(Kurtzer et al., 2017)

Ambientes portables

- Máquinas virtuales.
- Contenedores.

Máquinas virtuales



Proporcionan un ambiente completo, incluyendo dependencias de software, bibliotecas, datos que pueden ser encapsulados y ejecutados donde sea.

Problema

Introduce una sobrecarga computacional considerable debido al nivel requerido de virtualización para emular el sistema operativo y el kernel.

Contenedores



Características de virtualización ligera en el kernel de Linux hace posible construir virtualización más ligera.

Las características específicas del kernel son las que se ocupan de aislar procesos, en particular (Nemeth et al., 2018):

- **Namespaces** : aíslan los procesos del contenedor desde la perspectiva de las características del sistema operativo.
- **Control groups (cgroups)** limita el uso de los recursos del sistema y prioriza ciertos procesos sobre otros.

Apptainer

2015



SINGULARITYCE



Sylabs



2021



APPTAINER





Gregory Kurtzer



Rocky Linux™

Objetivos de Apptainer (Kurtzer et al., 2017)

- **Mobility of compute (Portabilidad)**
- **Reproducibilidad**
- **Soporte e integración con recursos tradicionales de HPC**

Mobility of compute

- Capacidad de definir, crear y mantener un workflow local que puede ser usado con confianza en diferentes hosts, sistemas operativos de Linux, proveedores de servicios en la nube, etc.
- Esencial para la *ciencia reproducible*.
- Apptainer usa un formato de imagen distribuible que encapsula el contenido completo del contenedor en un **archivo de imagen único** (archivo **.sif**).

Mobility of compute

- El archivo .sif es la representación completa de todos los archivos dentro del contenedor.
- El .sif encapsula el sistema operativo así como también todas las dependencias necesarias para ejecutar un workflow definido.
- Este archivo puede ser copiado, compartido, mejorado, además de seguir el estándar de permisos de UNIX.

Reproducibilidad

- Si se garantiza portabilidad, se garantiza reproducibilidad.
- Una vez que se ha definido el workflow dentro de contenedor, el archivo .sif puede usarse con confianza que el código dentro del contenedor no ha sido modificado.
- Apptainer utiliza un método de validación de integridad mediante un hash por SHA256 que garantiza que la imagen del contenedor que es distribuida no ha sido modificada.
- Cuando las investigadoras publican sus resultados, pueden distribuir el .sif de la imagen utilizada en el artículo así como su hash, permitiendo que alguien más pueda validar los resultados y confirmar la integridad del .sif descargado con el hash.

Soporte e integración con recursos tradicionales de HPC

Apptainer soporta de forma nativa tecnologías como :

- Infiniband y Lustre (comunicación en red de alta velocidad y baja latencia y sistema de archivos paralelo, respectivamente).
- SLURM, Torque, etc (Sistemas de administración de recursos).
- Aceleradores: GPUs, TPUs.
- Software tradicional de HPC: OpenMP, MPI.

Esto como resultado que Apptainer se ejecuta como cualquier otro comando del sistema.

Investigaciones académicas

- Desarrollan en ambiente local, escalan la ejecución en otra infraestructura.
- Publican resultados y destruyen el análisis completo y el workflow dentro de un .sif acompañado del hash, permitiendo la reproducción y validación de resultados.

Administración de sistemas

- El sistema de HPC es controlado y administrado por una administradora de sistemas y su equipo.
- Para mantener el sistema seguro, sólo a este grupo se le concede acceso root y control sobre el estado del sistema operativo.
- Los usuarios del sistema no tiene acceso como root.
- Si un usuario puede escalar a root (incluso dentro de un contenedor) en el sistema, el usuario puede potencialmente hacer cosas malas (deliberadamente o no).

Administración de sistemas

- Para mitigar estos riesgos, Apptainer no proporciona la capacidad de escalar permisos dentro de contenedor.
- Con Apptainer **si un usuario no tiene acceso a root en el sistema objetivo, tampoco puede escalar privilegios como root dentro del contenedor.**
- La sysadmin instruiría a las usuarias a desarrollar sus imágenes en un sistema donde tengan acceso a root para realizar operaciones de escritura en las imágenes, para posteriormente transferirlas al cluster para su ejecución a gran escala.

Eliminar redundancia en tecnologías de contenedores

- Es posible ejecutar imágenes de Docker en ambientes HPC con Apptainer.
- Por diseño, Apptainer ha sido desarrollado para trabajar sin problemas con Docker.

¿Por qué no sólo usar Docker en ambientes HPC?

Problemas de seguridad

- Para cada contenedor que Docker ejecuta, el proceso del contenedor es extendido como hijo del demonio de Docker que es propiedad de root.
- Como la usuaria es capaz de interactuar y controlar el demonio de Docker, teóricamente es posible obligar al proceso del demonio a otorgar permisos privilegiados a la usuaria¹.

¹Uno de los principales retos en los centros HPC dedicados a la investigación es permitir a las usuarias ejecutar código arbitrario manteniendo simultáneamente que el sistema no se encuentra comprometido por código malicioso (intencionalmente malicioso o no).

Workflow de Apptainer

- **Build.** Construye el contenedor de Apptainer en un sistema local donde tengas acceso a root o sudo.
- **Transfer.** Transfiere el contenedor a un sistema HPC donde quieras ejecutarlo.
- **Run.** Ejecuta el contenedor en el sistema HPC.

References I

- Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459.
- Nemeth, E., Snyder, G., Hein, T. R., Whaley, B., and Mackin, D. (2018). Unix and linux system administration handbook. *USENIX Open Access Policy*, 59.