FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

EPN GRID Networking

Dirk Hutter (FIAS)

David Rohr (CERN)

Introduction

- The overall goal is to run ALICE GRID jobs on the ALICE online farm on nodes which are not used for data taking
- Other nodes of the farm are used for data taking in parallel to GRID jobs
- Results from the jobs need to be copied to grid sides outside of CERN.
- The jobs run in singularity/appteiner containers and must access the GPUs
- To make data transfer possible we agreed on the following solutions:
  - Allow a set of public IPs on an isolated IB partition of the EPN cluster
  - Assign these public IPs only to the containers. The IP must not be assigned to the host, i.e., containers have an isolated network
  - The data is routed via the link between CR0 and CERN CC (not via the GPN)

Current „Async“ Setup

- The execution flow is:
  - A grid jobs is scheduled on the EPNs via slurm and starts a grid agent
  - This grid agent then starts an apptainer container running the actual processing
  - This container needs to communicate to the outside
- For efficient GPU usage the two NUMA domains of the EPNs are scheduled independently via slurm, i.e. there are usually two concurrent slurm jobs per EPN (and thus also two grid agents and two containers)
- Slurm does the needed NUMA control
- The slurm jobs run with cgroups support, i.e., slum job_containers
- The EPNs are connected via an InfiniBand network, which connected to the CERN LCG ETH network via IB-ETH gateways

Proposal

- Fully isolated network from GRID containers on to IT router in CR0
  - On the node: run the GRID containers in an isolated network namespace
  - On the IB fabric: create an isolated partition for the new subnet
- Proof-of-concept available
  - Create virtual IB HCAs via SR-IOV and assign the ports to a separate IB partition
  - Move the vHCA to the containers namespace before configuring any public IP and route
  - Based on: local apptainer install (with setuid support) + CNI ib-sriov-cni plugin + configuration files
- We have already tested successfully to start the default GRID container with our local proof-of-concept apptainer installation and the network isolation.
  - No modification of default grid container on CVMFS needed, we use a local apptainer installation
  - Newest version of GRID agents supports explicitly setting apptainer installation to use
  - No sudo - rights / etc. for GRID user (alicesgm). Privileges for moving HCAs to network namespace provided by apptainer suid wrapper from local apptainer installation

Suggested Separation

- EPN team
  - The EPN team provides the vHCAs, IB partition configuration and routing on the IB-ETH gateway
  - The EPN team will take care of installation of configurations, plugins and packages, e.g., apptainer installation, CNI ib-sriov-cni plugin and the apptainer-network-config files on the nodes
- PDP team
  - Will take care of the wrapper to start the local container installation with the right network-config
  - The proper vobox configuration
- GRID team
  - Provides the possibility to run a custom apptainer installation (already available)
  - Takes care of containers (network) security, e.g., open ports, firewalls, …

- Bonus

Isolation on the IB fabric

- Add a dedicated IB partitions which is used for the IPoIB subnet with public IPs.
- The IB-ETH gateway joins this partition and the respective IPoIB subnet.
- The IB-ETH gateway is set as gateway of this subnet
  - Any IPs beside the subnet configured in the IB gateway will not be route to the outside
- A host can joint this partition in one of two ways:
  - The IPoIB (parent) interface always joins the default partition (default = first pkey in the pkey table)
  - A child interface can be created for any other configured partition key
  - A virtualized HCA (SR-IOV) has a virtual port with it's own port GUID. We can freely assign a different pkey table (also for the default pkey).