DE LA RECHERCHE À L'INDUSTRIE

Virtualization on high performance compute clusters with pcocc

www.cea.fr

Teratec - June 2016

Virtualization with pcocc
Agenda

Motivation

Presentation of the tool

Performance and results

Motivation

Virtualization with pcocc
Motivation

Compute clusters are getting more mainstream

- More and more diverse user communities
  - Heterogeneous requirements
  - New software stacks
  - Application development
- Computing centers are expected to provide new services
  - Offer on-site pre/post treatment tools
  - Allow hosting user services
- New tools are required to address these requirements

ESGF
Earth System Grid Federation

ccrt

Virtualization with pcocc
Motivation

Virtualization and HPC are no longer incompatible

- Hardware virtualization improvements
  - Minimal performance overhead
  - Support for HPC interconnects (Infiniband SR-IOV)

- More and more examples of real world uses
  - On-demand HPC clusters using cloud providers
    - Elasticluster / MIT Starcluster / CycleComputing
  - Grid-computing at CERN
    - CernVM virtual appliances
  - Comet (SDSC) « The world's first virtualized supercomputer »
    - « VM jobs scheduled just like batch jobs »
    - « VMs will be easy on-ramp for new users/communities, including low porting time »
  - Cori (NERSC)
    - « User-Defined Images : Enables users to accompany applications with portable, customized OS environments »
    - Developed a container based solution: Shifter

Virtualization with pcocc
Motivation

Expected benefits of virtualization at CEA

- Provide the users with full control over their software environment
  - Allow satisfying all the dependencies of an application down to the OS
  - Applications can be packaged with their software stack in an image
  - Avoids reproductibility issues due to subtle software environment changes

- Enable new uses of compute resources
  - Facilitate the work of developers of scientific applications
    - Ability to perform tests in various software environments
    - Continuous integration
  - Allow test and development of system tools
    - Develop and test system software at large scale
    - Avoids having to setup dedicated resources
    - Work can be performed without administrative privileges
      - Internships

=> Required a tool to easily deploy VMs on our existing HPC clusters

Presentation of the tool

Virtualization with pcocc
Presentation of the tool

Private Cloud on a Compute Cluster

- Tool to easily deploy virtualized workloads on an existing compute cluster
  - Allow using the cluster as a kind of « private cloud »
    - Instantiate virtual clusters in the same way as jobs
    - Full administrative privileges and control over VM image
  - Resources are managed by SLURM
    - Usual semantics of resource allocation
    - One task = One VM
    - VMs are automatically sized depending on the underlying resources (CPU/memory)
  - Each virtual cluster has its own private isolated networks
    - VMs are interconnected with Ethernet and/or Infiniband
  - Integration to the native cluster environment
    - Reverse NAT for SSH access
    - Host NFS/Lustre filesystems can be reexported via 9P

Virtualization with pcocc
Presentation of the tool

Usage overview

- List available VM templates
  - ***pcocc template list***

```
NAME                DESCRIPTION                          RESOURCES      IMAGE
----                -----------                          ---------      -----
compute             Centos7 based compute node           ib             /path/to/compute-image
master              Centos7 based master node            ib             /path/to/master-image
ci-centos7          Vanilla CentOS7 cloud-init image     eth            /path/to/cloud-image
```

- Allocate 128 8-cores VMs from the 'compute' template and 1 from the 'master' template
  - ***pcocc alloc -c 8 master:1,compute:128***
  - Each VM disk is an ephemeral CoW image based on the selected template

- Connect to the first vm of a virtual cluster via ssh
  - ***pcocc ssh [-j <jobid>] root@vm0***

- Creating templates
  - Save a new template or a new revision from a running VM
    - ***pcocc save [-j <jobid>] [-d newimage] vm5***
  - Supports cloud-init (tool to customize vanilla images from various distributions)
  - Import any qcow2 file in the user storage spaces

Virtualization with pcocc
Presentation of the tool

SLURM integration

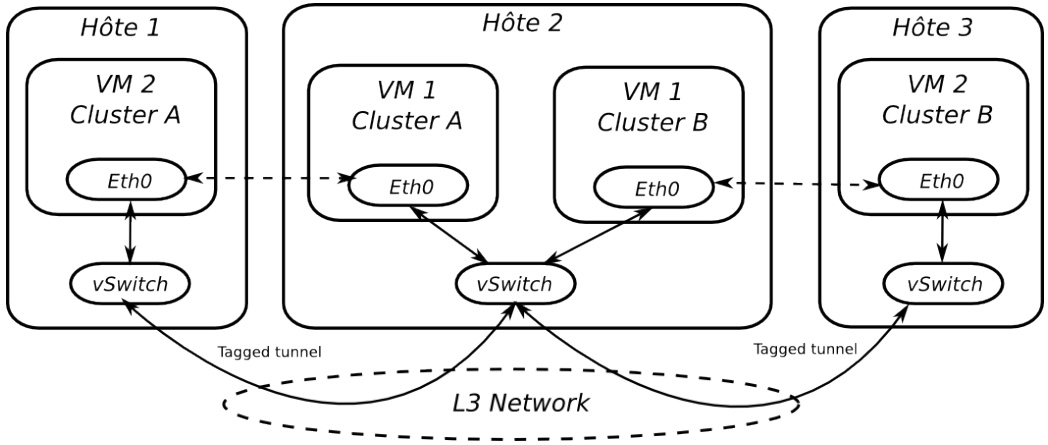
- SLURM spank plugin + prolog/epilog scripts
  - Performs all privileged operations required to launch VMs
    - Creates and configures TAP devices
    - Configures VFIO for SRIOV passthrough
    - Sets up iptables and OpenFlow rules
    - Assigns Infiniband pkeys
  - Qemu is launched as a regular SLURM task
    - Uses network resources created during prolog
    - VM defined to closely match underlying host resources
      - Virtual CPU and memory pinning taking NUMA nodes into account

Virtualization with pcocc
Presentation of the tool

Private Ethernet networks

- Universally supported interconnexion network
- Easily virtualizable by software
- IP tunnels are created between compute nodes
  - GRE encapsulation of VM Ethernet packets
  - Packets may be relayed over any L3 layer network (IpoIB)
  - Implementation based on OpenVswitch
  - ~350MB/s throughput over QDR Infiniband

OPEN vSWITCH
An Open Virtual Switch



Virtualization with pcocc
Presentation of the tool

Private Infiniband networks

- Exposing Infiniband to VMs is required for tightly coupled parallel applications
  - « OS-bypass » makes efficient software virtualization difficult
  - Direct access to the hardware is required

- Leveraging Infiniband SR-IOV support
  - Hardware multiplexing of a device into multiple virtual functions
    - A physical function remains in charge of the configuration of the device
      - Managing virtual functions, assigning GUIDs and PKeys
    - Virtual functions are restricted to data transfer
      - No access to QP0, QP1 is para-virtualized

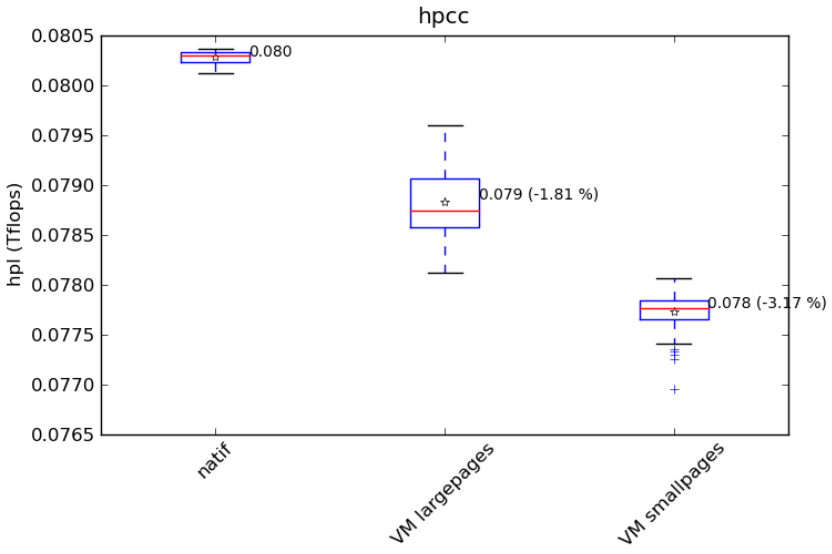
- Ensuring Isolation
  - A PKey (Infiniband's equivalent to a VLAN) is allocated for each virtual cluster
    - OpenSM is dynamically reconfigured to associate PKeys to host nodes
    - Virtual functions are restricted to using the allocated Pkey
  - An IOMMU ensures that each VF can only access it's VM memory
    - Leverages the kernel VFIO module

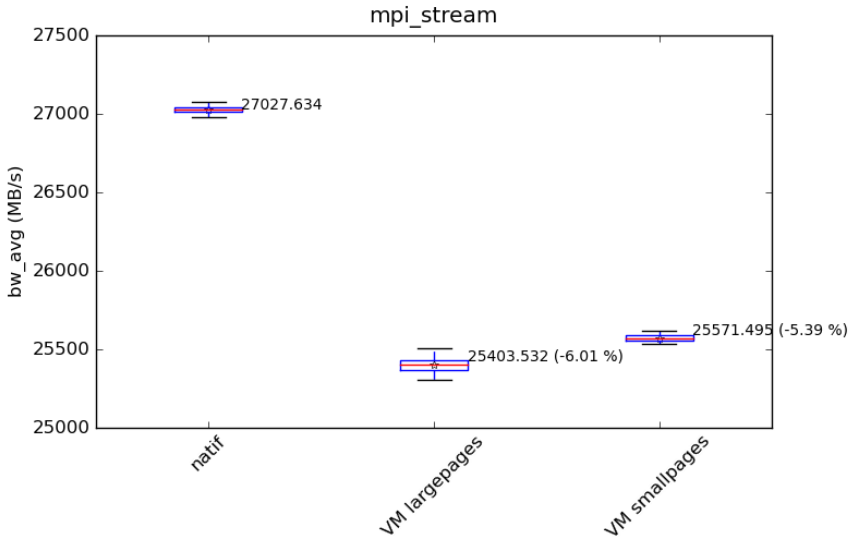
Performance and results

Performance results

- In-house applicative benchmarks suite
  - Automated batch submission of benchmarks
  - Ran on our R&D cluster (inti)
    - 128 Bull B500 bi-Nehalem nodes (2x4 cores) 2.8Ghz
    - 36 Bull B720 bi-Haswell nodes (2x16 cores) 2,3Ghz

- Low impact of virtualisation on single compute node performance

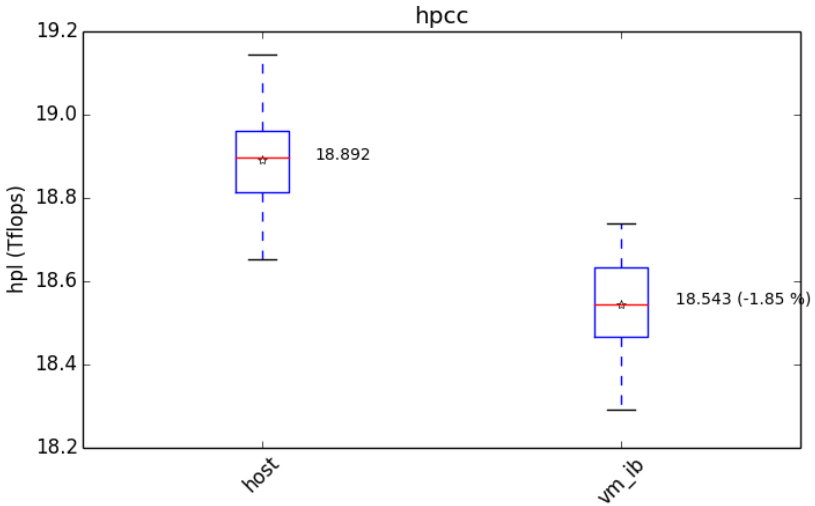


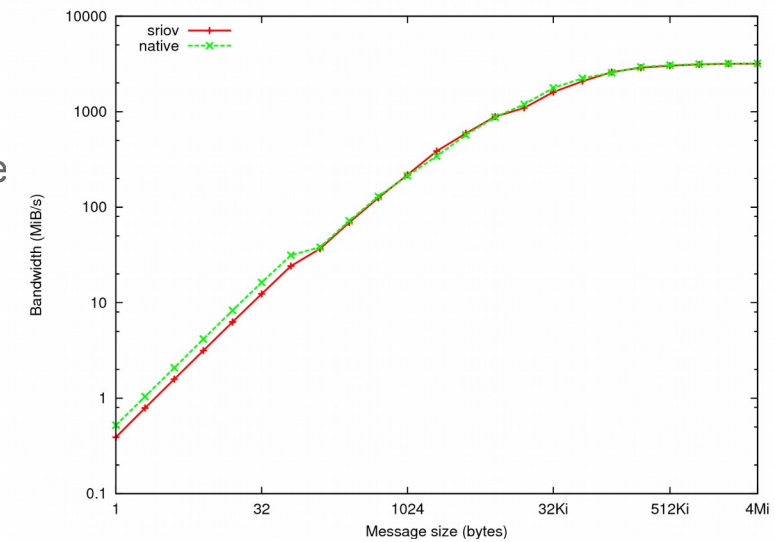


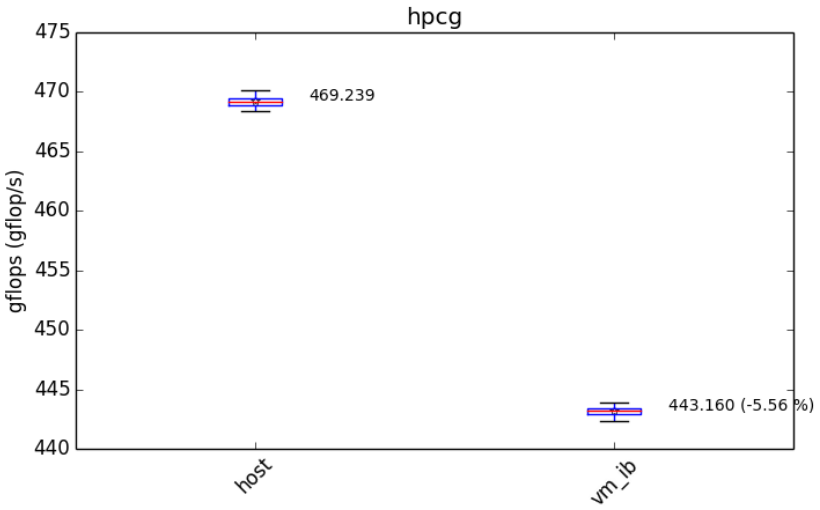
Virtualization with pcocc
Performance

Performance results

- First parallel benchmarks with virtualized Infiniband
  - Bandwidth / Latency close to native performance
  - Good results on first benchmarks (1024 cores)







Virtualization with pcocc
Results

First successful uses

- Puppet internship
  - Performance evaluation of our configuration management tools
    - At the scale of a cluster
    - In various environments
      - CentOS 6 and 7
      - Various ruby and passenger configurations
  - Would have been hard without easy access to virtualization
    - Isolating a large number a physical nodes
    - Tedious setup process
  - Pcocc allowed to perform tests up to 512 VMs on 1024 cores
    - Each test ran in a 1024 core batch job for a few minutes
- Five other internships are underway (on Lustre, networking technologies, ...)

Virtualization with pcocc
Results

First successful uses

- Non-regression testing of system-level software
  - Nfs-ganesha : user-mode file server
  - Jenkins is plugged into SLURM and submits pcocc jobs
  - A virtual cluster (NFS server and client) is instanciated for each test

- Preparing for the arrival of new generations of clusters
  - Validating the behaviour of critical applications in new software environments

Future work

- Hardware virtualization overhead is acceptable for medium-size HPC jobs
  - Evaluation at larger scale to be performed
- Virtualization allows us to leverage our clusters for new tasks
  - Less need for dedicated test clusters
- We will pursue our investment into virtualization
  - Allow users to host their own services
  - Evaluate OS level virtualization (containers)

Thank you for your attention

Questions ?

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM Ile-de-France | Bruyères-le-Châtel 91297 Arpajon Cedex
T. +33 (0)1 69 26 40 00 |
Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019

DAM/DIF
DSSI
SISR