Using Docker in High Performance Computing in OpenPOWER Environment

Zhaohui Ding, Senior Product Architect
Sam Sanjabi, Advisory Software Engineer
IBM Platform Computing
Agenda

- What is LSF?
- What is Docker?
- LSF & Docker
- Next step
- Q & A
What is LSF?

I need access to 32 RHEL 6.1 host computers connected via infiniband, licensed to run LS-DYNA with Platform MPI and having at least 4 GB of RAM per node for four hours.

USER 1
Project A

USER 2
Project B

USER 3
Project C
What is LSF?

```bash
user2% bsub -n 32 -P ProjectB \
-R "rhel61 && mem>4000 && ibswitch && lsdyna" \
-app lsdyna -k "chkpointdir method=lsdyna" \
-R "rusage[mem=3500:lsdyna=32]" \
dyna_simulation.sh
```
LSF Architecture

LSF supports multiple scheduling policies that can be combined in flexible ways.

IBM PLATFORM LSF SCHEDULER

- Preemption
- Fairshare
- Allocation Units
- Resource Reservation
- Advanced Reservation
- Advance Parallel
- User-Defined
- Multicluster

PLUG-IN SCHEDULERS

- Local Clusters
- Remote Clusters

Join the conversation at #OpenPOWERSummit
What is **Docker**?

- A lightweight container technology built on top of Linux container, cgroup and AUFS

- Hot topic in “cloud & big data” to develop, ship and run applications anywhere and solve “dependency hell” challenges

Join the conversation at #OpenPOWERSummit
Docker Architecture

- The Docker Daemon
- The Docker Client
- Inside Docker
  - Docker Images
  - Docker Registries
  - Docker Containers

Join the conversation at #OpenPOWERSummit

http://docs.docker.com/article-img/architecture.svg
OpenPower

IBM Stack
Research And Innovation

IBM
Mellanox
OpenPower
Open Innovation

Google

NVIDIA

TYAN

IBM POWER8

Google
IDC System,
Software Stack

CAPI / PCI G3
Directly Integrate
Partner IP into Power8
Chip protocols

Mellanox
World class
Networking and
Switching

NVidia
World class
GPU Compute
Capabilities

Join the conversation at #OpenPOWERSummit
Containers in HPC

- Not a new idea ...
  - Workload resource isolation
  - Process tracking
  - Job controlling
  - Checkpoint/restart, live migration
  - ...

- Examples:
  - IBM AIX WLM
  - Linux control group
  - Virtual machine
  - ...

Join the conversation at #OpenPOWERSummit
Docker in HPC

Join the conversation at #OpenPOWERSummit
Potential Benefits

- Resource guarantee and performance isolation
  - Docker leverages Linux control groups

- Application encapsulation and cloud mobility
  - Docker makes it easy

- Application lifecycle management
  - Different versions of applications coexisting in the same environment

- Consistency, repeatability and compliance
  - Docker provides a way to run application in a known predefined environment.

- Lightweight, fast and transparent
  - Identical application performance on bare metal systems and fast management operations
How to use it?

[lsfadmin@ppcrhel7e bin]# lshosts
HOST_NAME   type  model  cpuf  ncpus  maxmem  maxswp  server  RESOURCES
ppcrhel7e   LINUXPP  POWER8  250.0  19  128G  3.9G  Yes (mg docker)

[lsfadmin@ppcrhel7e ~]$ bsub -Is -a "docker(apollos/ubuntu-ppcle:trusty)" /bin/sh
Job <105> is submitted to default queue <interactive>.
<<Waiting for dispatch ...>>
<<Starting on ppcrhel7e>>
$ echo $LSB_JOBID
105
$ hostname
lsf_docker-105
$ cat /etc/*release*
DISTRIBUT_ID=Ubuntu
DISTRIBUT_RELEASE=14.04
DISTRIBUT_CODENAME=trusty
DISTRIBUT_DESCRIPTION="Ubuntu 14.04.1 LTS"

[root@ppcrhel7e bin]# lsrun -m ppcrhel7e docker ps
CONTAINER ID        IMAGE                         COMMAND             CREATED             STATUS              PORTS
f006a5a8dd7c      apollos/ubuntu-ppcle:trusty   "/bin/bash"         2 minutes ago       Up 2 minutes

Cluster name and LSF JobID
Docker image name
Docker Env deployed

Join the conversation at #OpenPOWERSummit
How it works?

- Leverage existing IBM Platform LSF plugins mechanism (esub, job starter, job control etc)
Resource Requirement & Enforce

- **CPU Shares**
  - Translate IBM Platform LSF total number of slots to relative cpu share on docker.
  - `bsub -n <num> -> docker run --cpu <num>`

- **Memory Fencing**
  - Enable IBM Platform LSF cgroup feature and translate –M/-v to docker memory option
  - `bsub -M/-v -> docker run --memory`

- **Affinity & NUMA aware scheduling**
  - IBM Platform LSF selects which hosts and cores/sockets and pass information to docker to enforce it.
  - `bsub -n 4 -R “span[hosts=1] affinity[core]” -> docker run --cpuset`
Next Steps

- Tighter integration with IBM Platform LSF job resource collection framework
  - Leverage built-in runtime IBM Platform LSF resource collection and reporting mechanism instead of using bpost

- Docker image aware scheduling
  - Docker image can be big and it takes time to download and install it.
  - Prefer to go to hosts with image pre-installed

- Local image management
  - Image should be cleaned up periodically when not required to save disk space
Please Download and Try It Out


White paper: https://ibm.biz/BdFzPY
Q & A

Thank You