

Secure Multiparty Computation Sprint 4

Developer | Hasnain Abdur Rehman | hasnain@bu.edu
Developer | Pierre-François Wolfe | pwolfe@bu.edu
Developer | Samyak Jain | samyakj@bu.edu
Developer | Suli Hu | sulihu@bu.edu
Developer | Yufeng Lin | yflin@bu.edu
Mentor/Client | John Liagouris | liagos@bu.edu
Mentor/Client | Vasiliki Kalavri | vkalavri@bu.edu
Subject-Matter Expert | Mayank Varia | varia@bu.edu



Presentation Outline

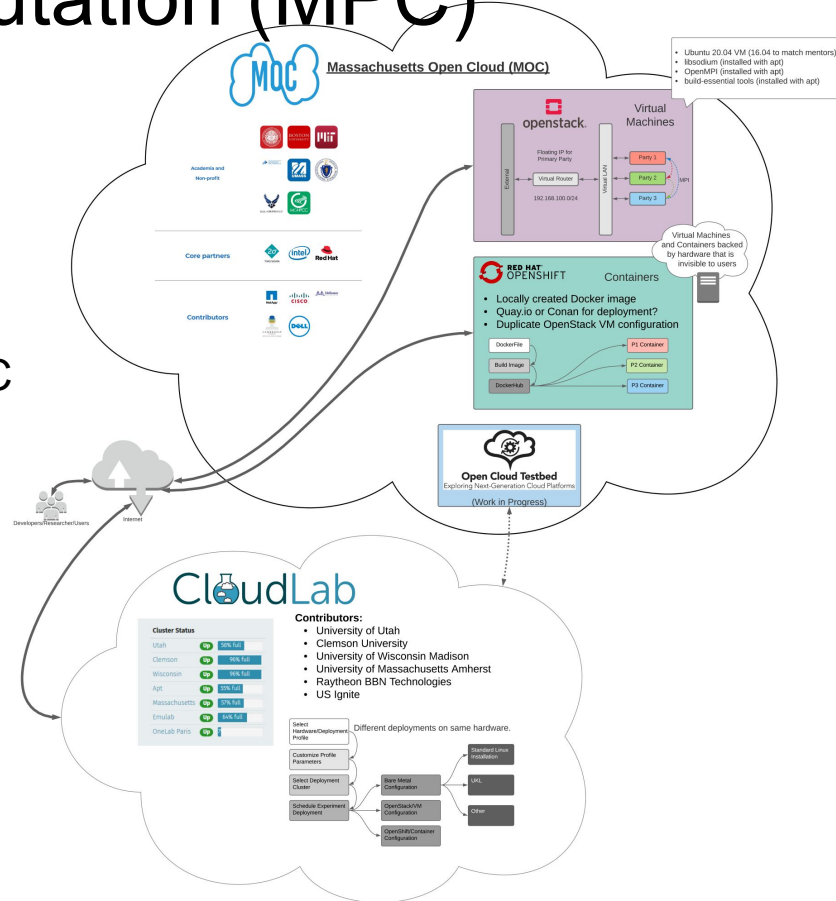
- Project Recap
- Project Goals & Sprint 4 Stories/Tasks
- Work Accomplished & Information Learned
 - Bare-Metal → CloudLab Testing
 - Containers → Docker/OpenShift
- Project Organization Assessment (Burndown)
- Sprint 5 goals (Mentor priorities)



Recap of Multi-Party Computation (MPC)


- MPC enables...
 - Shared Computation on Private Data
 - Protects the Privacy of Data
 - Mutually Agreed Computation
- Our mentors...
 - Are using three party Secret Sharing MPC
 - Perform Relational Queries with MPC
 - Keep all parts secure vs. splitting into secure and insecure steps
- Our mission...
 - Profile this new MPC library
 - Identify bottlenecks
 - Compare deployment scenarios and find the best performance

Boston University CS & ECE



Project Goals & Sprint 4 Stories/Tasks

- Presentation
 - Finalize SmartNIC presentation
- CloudLab
 - Geni-lib automation
 - Bare metal testing
- Containers
 - Debug OpenMPI on Docker
 - Docker/Docker-compose work
 - Status with OpenShift (kompose)

▼ Sprint 4	
29 Oct 2020-12 Nov 2020	60 closed 88 total
	
#117 As a student in the CC course, I want prepare and deliver a practice presentation about the selected article to be ready for the class presentation. 🕒	40
#11 As a team member, I want to build a containerized MPC environment on OpenShift, like one on the VMs.	12
#162 As a team member, I want to benchmark the results obtained by running the codebase on multi bare-metal nodes	16
#211 As a team member, I want to create a demo summarizing accomplishments in order to show progress to the clients	20



Bare-Metal Testing

CloudLab, geni-lib scripts, data collection, ...

CloudLab: Geni-lib Automation

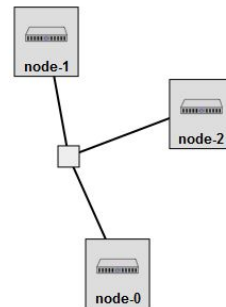
- LAN vs Link
- Static IP
- Setup
 - Dependencies
 - MPC Code duplicate
 - geni.rspec.pg.Install
- Remaining Issue:
 - geni.rspec.pg.Execute

Boston University CS & ECE

```

1  """ubuntu baremetal ring of nodes"""
2
3  #
4  # NOTE: This code was machine converted. An actual human would not
5  #       write code like this!
6  #
7
8  # Import the Portal object.
9  import geni.portal as portal
10 # Import the ProtoGENI library.
11 import geni.rspec.pg as pg
12 # Import the Emulab specific extensions.
13 import geni.rspec.emulab as emulab
14
15 pc = portal.Context()
16
17 pc.defineParameter("node_type", "Hardware Type",
18                  portal.ParameterType.NODETYPE, "any")
19 pc.defineParameter("node_count", "Number of Machines",
20                  portal.ParameterType.INTEGER, 3)
21
22 params = pc.bindParameters()
23
24 request = portal.context.makeRequestRSpec()
25
26 node = []
27 link = []
28
29 # Create selected number of nodes
30 for i in range(params.node_count):
31     node.append(request.RawPC('node-%d' % i))
32     node[-1].disk_image = 'urn:publicid:IDN+emulab.net+image+emulab-ops:UBUNTU16-64-STD'
33     node[-1].hardware_type = params.node_type
34
35 # Create a LAN for all the connections
36 lan = request.LAN("lan")
37
38 # Create a link between each of the nodes to make a ring
39 for i in range(params.node_count):
40     iface = node[i].addInterface("if1")
41     iface.component_id = "eth1"
42     iface.addAddress(pg.IPv4Address("192.168.1."+str(i+1), "255.255.255.0"))
43     lan.addInterface(iface)
44
45 # Install and execute scripts on each node
46 for i in range(params.node_count):
47     node[i].addService(pg.Install(url="https://www.dropbox.com/s/7t91cf0ugt66ypl/cloudlab_setup.tar.gz", path="/home/mpc"))
48     node[i].addService(pg.Execute(shell="bash", command="/home/mpc/setup.sh"))
49
50 # Print the generated rspec
51 pc.printRequestRSpec(request)

```



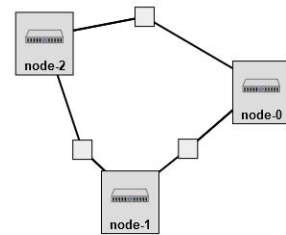
CloudLab: Geni-lib Further Tweaks

- Modifications
 - Link multiplexing
 - Best Effort
- Issues with Ring Topology
 - Couldn't map to hardware
- Current Status: Still WIP

```

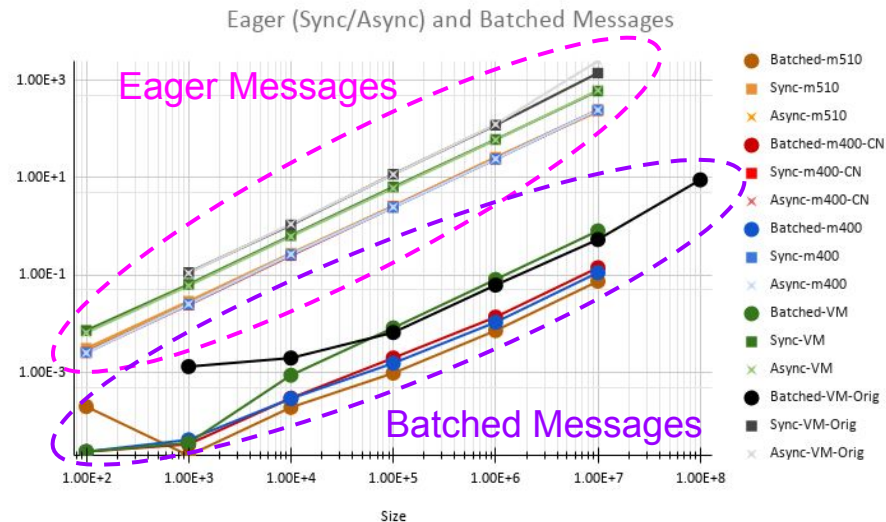
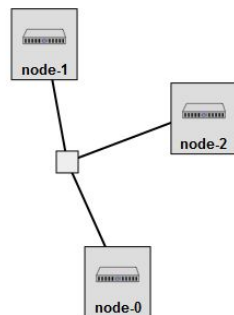
1  """ubuntu baremetal ring of nodes"""
2
3  #
4  # NOTE: This code was machine converted. An actual human would not
5  #       write code like this!
6  #
7
8  # Import the Portal object.
9  import geni.portal as portal
10 # Import the ProtoGENI library.
11 import geni.rspec.pg as pg
12 # Import the Emulab specific extensions.
13 import geni.rspec.emulab as emulab
14
15 pc = portal.Context()
16
17 pc.defineParameter("node_type", "Hardware Type",
18                  portal.ParameterType.NODETYPE, "any")
19 pc.defineParameter("node_count", "Number of Machines",
20                  portal.ParameterType.INTEGER, 3)
21
22 params = pc.bindParameters()
23
24 request = portal.context.makeRequestRSpec()
25
26 node = []
27 link = []
28
29 # Create selected number of nodes
30 for i in range(params.node_count):
31     node.append(request.RawPC(["node-%d" % i]))
32     node[i].disk_image = 'urn:publicid:IDN+emulab.net+image+emulab-ops:UBUNTU16-64-STD'
33     node[i].hardware_type = params.node_type
34
35 # Create a LAN for all the connections
36 #lan = request.LAN("lan")
37
38 # Create a link between each of the nodes to make a ring
39 for i in range(params.node_count):
40     iface = node[i].addInterface("if1")
41     # iface.component_id = "eth1"
42     # iface.addAddress(pg.IPv4Address("192.168.1."+str(i+1), "255.255.255.0"))
43     # lan.addInterface(iface)
44
45 # Create two links between them
46 link1 = request.Link(members = [node[0], node[1]])
47 iface1 = node[0].addInterface("if1")
48 iface1.component_id = "eth1"
49 iface1.addAddress(pg.IPv4Address("192.168.1."+str(1), "255.255.255.0"))
50 link1.addInterface(iface1)
51
52 link2 = request.Link(members = [node[1], node[2]])
53 iface2 = node[1].addInterface("if2")
54 iface2.component_id = "eth1"
55 iface2.addAddress(pg.IPv4Address("192.168.1."+str(2), "255.255.255.0"))
56 link2.addInterface(iface2)
57
58 link3 = request.Link(members = [node[2], node[0]])
59 iface3 = node[2].addInterface("if3")
60 iface3.component_id = "eth1"
61 iface3.addAddress(pg.IPv4Address("192.168.1."+str(3), "255.255.255.0"))
62 link3.addInterface(iface3)
63
64 # Turn on link multiplexing. Note that this also turns on vlan encapsulation
65 # You have to set this both links.
66 link1.link_multiplexing = True
67 link2.link_multiplexing = True
68 link3.link_multiplexing = True
69
70 # But the resource mapper is going to try to prevent the two links from oversubscribing
71 # the physical link. For example, trying to create two 1Gb multiplexed links on top of a 1Gb
72 # physical link. Sometimes this is the correct behaviour. But if not, do this to turn
73 # off the checks.
74 link1.best_effort = True
75 link2.best_effort = True
76 link3.best_effort = True
77
78 # Install and execute scripts on each node
79 for i in range(params.node_count):
80     node[i].addService(pg.Install(url="https://www.dropbox.com/s/791cf0ugt66ypl/cloudlab_setup.tar.gz", path="/home/mpc"))
81     node[i].addService(pg.Execute(shell="bash", command="/home/mpc/setup.sh"))
82
83 # Print the generated rspec
84 pc.printRequestRSpec(request)

```

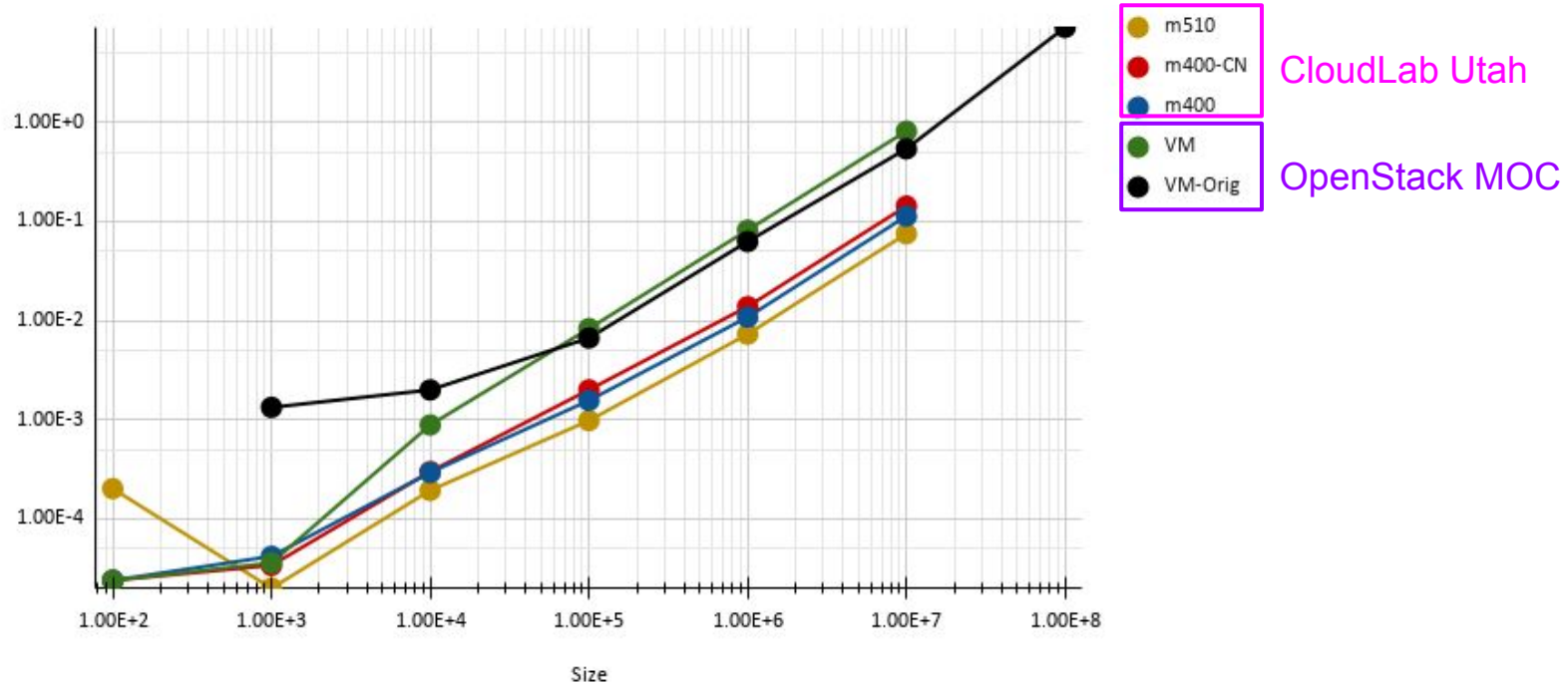


CloudLab Testing

- Shared Control Network
 - Hardware
 - m400 (ARM)
- LAN topology
 - Hardware
 - m510 (x86_64)
 - m400 (ARM)

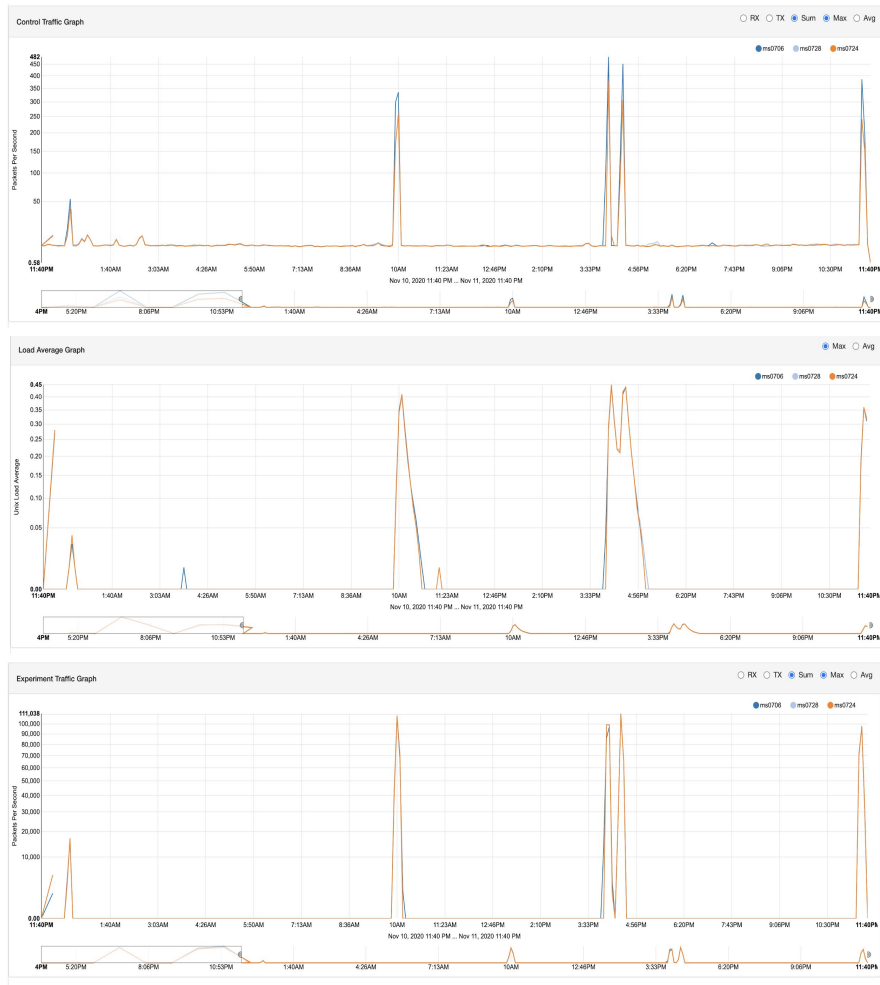


Batched Messages - Different Deployments



Observations

- News tests are using new Ubuntu image (14.04.1 → 16.04.01)
- When running 10M size test, performance varies might because of the time period. Test gets stuck in peak-hours and function normally in early morning EST.
- Experiments of smaller sizes are not restricted by time period.



Working with Containers

Docker, docker-compose, OpenShift, ...

MPC in Docker Container Debugging...

- Running tests.sh → OK
- Running exp-exchange
 - Issue with size greater than 505... ex: with 1000
 - Cryptic message...
- Determine source...
 - Some clues but overall meaning still unclear

```
mpirun -np 3 exp-exchange1000  
# which produces the following:  
root@ebd1c7f24dfe:~/experiments#
```

```
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00045] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1  
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1  
BATCHED 1000 0.00005  
SYNC 1000 0.00047  
ASync 1000 0.00042
```

Size changes value

[Hostname:PID]

OpenMPI Debugging Continued...

```
$12 = {MPI_SOURCE = 0, MPI_TAG = 193, MPI_ERROR = 0, _cancelled = 0, _count = 8000}
(gdb) p result1
$13 0
(gdb) p status2
$14 = {MPI_SOURCE = 0, MPI_TAG = 193, MPI_ERROR = 0, _cancelled = 0, _count = 8000}
(gdb) p result2
$15 0
(gdb) l
220     }
221     else { //P3
222         result1 = MPI_Recv(r1s1, ROWS, MPI_LONG_LONG, 0, SHARE_TAG, MPI_COMM_WORLD, &status1);
223         result2 = MPI_Recv(r1s2, ROWS, MPI_LONG_LONG, 0, SHARE_TAG, MPI_COMM_WORLD, &status2);
224     }
225 }
```

MPI_Recv → No error directly returned...

```
mpirun -np 3 exp-exchange 1000
# which produces the following:
root@ebd1c7f24dfe:~/experiments#
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00045] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00047] Read -1, expected 8000, errno = 1
[ebd1c7f24dfe:00046] Read -1, expected 8000, errno = 1
BATCHED 1000 0.00005
SYNC 1000 0.00047
ASYNC 1000 0.00042
```

Looking for matching message pattern on github...

```
68 #if OPAL_BTL_SM_HAVE_CMA
69 int mca_btl_sm_get_cma (mca_btl_base_module_t *btl, mca_btl_base_endpoint_t *endpoint, void *local_address,
70                        uint64_t remote_address, mca_btl_base_registration_handle_t *local_handle,
71                        mca_btl_base_registration_handle_t *remote_handle, size_t size, int flags,
72                        int order, mca_btl_base_rdma_completion_fn_t cbfunc, void *cbcontext, void *cbdata)
73 {
74     struct iovec src_iov = {.iov_base = (void *) (intptr_t) remote_address, .iov_len = size};
75     struct iovec dst_iov = {.iov_base = local_address, .iov_len = size};
76     ssize_t ret;
77
78     /*
79      * According to the man page :
80      * "On success, process_vm_readv() returns the number of bytes read and
81      * process_vm_writev() returns the number of bytes written. This return
82      * value may be less than the total number of requested bytes, if a
83      * partial read/write occurred. (Partial transfers apply at the
84      * granularity of iovec elements. These system calls won't perform a
85      * partial transfer that splits a single iovec element.)".
86      * So since we use a single iovec element, the returned size should either
87      * be 0 or size, and the do loop should not be needed here.
88      * We tried on various Linux kernels with size > 2 GB, and surprisingly,
89      * the returned value is always 0x7ffff000 (fwiw, it happens to be the size
90      * of the larger number of pages that fits a signed 32 bits integer).
91      * We do not know whether this is a bug from the kernel, the libc or even
92      * the man page, but for the time being, we do as is process_vm_readv() could
93      * return any value.
94      */
95     do {
96         ret = process_vm_readv (endpoint->segment_data.other.seg_ds->seg_cpuid, &dst_iov, 1, &src_iov, 1, 0);
97         if (0 > ret) {
98             opal_output(0, "Read %ld, expected %lu, errno = %d\n", (long)ret, (unsigned long)size, errno);
99             return OPAL_ERROR;
100         }
101         src_iov.iov_base = (void *) ((char *) src_iov.iov_base + ret);
102         src_iov.iov_len -= ret;
103         dst_iov.iov_base = (void *) ((char *) dst_iov.iov_base + ret);
104         dst_iov.iov_len -= ret;
105     } while (0 < src_iov.iov_len);
106
107     /* always call the callback function */
108     cbfunc (btl, endpoint, local_address, local_handle, cbcontext, cbdata, OPAL_SUCCESS);
109
110     return OPAL_SUCCESS;
111 }
112 #endif
```

OpenMPI Debugging Resolution

- Issue for messages larger than ~1k
- Shared memory Byte Transport Layer (BTL)
 - “Sm” was the original version
 - “Vader” is the current version
- CMA (Cross Memory Attach)
 - Kernel support required for “Zero copy” mechanism
- Bypass with parameter
 - `mpirun --mca btl_vader_single_copy_mechanism none -np 3 exp-exchange 1000`

Boston University CS & ECE

disable CMA in vader #3270



hunsan opened this issue on Apr 2, 2017 · 9 comments



hunsan commented on Apr 2, 2017

Hello all,

We experienced problems with Open MPI when communication larger messages, for example with `MPI_Gather`, `MPI_Allgather`, etc. Large means messages larger than 1k (for very small messages the problem did not occur).

We received error messages like this

`Read -1, expected 8000, errno = 38`

I suspected that there is something wrong with the CMA support.

And indeed, until Open MPI 2.0.2 the configure script would have the following result

```
#define OPAL_BTL_SM_HAVE_CMA 0
```

Now, the new CMA detection method in 2.1.0 leads to

```
#define OPAL_BTL_SM_HAVE_CMA 1
```

That wouldn't be a problem if we could disable CMA in vader during configure, but it does not seem to be possible. So, `--with-cma=no` or `--without-cma` will have no effect and we will end up with

```
#define OPAL_BTL_SM_HAVE_CMA 1
```

Currently, we can set

`OMPI_MCA_btl_vader_single_copy_mechanism=none`

and the error messages will not show up.

It would be great if we could manually disable CMA support during configure.

Thank you

Working with Docker and docker-compose

- `docker-compose.yml`
 - Define topology
 - Multiple “services” each a container based on Dockerfile
 - Virtual network for communication
- Build Configuration
 - `docker-compose build`
- Launch Three Containers
 - `docker-compose up`
- Dockerfile
 - Install dependencies
 - Configure non-root user
 - Configure ssh
 - Build MPC code
- Build Image
 - `docker build -t mpc .`
- Launch Single Container
 - `docker run --name mpc -it mpc /bin/bash`

Container Configuration Hierarchy

docker-compose

Three
containers

Network
Config.

```

1 version: '2'
2
3 services:
4   party-0:
5     hostname: party-0
6     build:
7       dockerfile: Dockerfile
8       context: .
9     tty: true
10    networks:
11      mpc_net:
12        ipv4_address: 192.168.1.11
13    ports:
14      - "22"
15   party-1:
16     hostname: party-1
17     build:
18       dockerfile: Dockerfile
19       context: .
20     tty: true
21    networks:
22      mpc_net:
23        ipv4_address: 192.168.1.12
24    ports:
25      - "22"
26   party-2:
27     hostname: party-2
28     build:
29       dockerfile: Dockerfile
30       context: .
31     tty: true
32    networks:
33      mpc_net:
34        ipv4_address: 192.168.1.13
35    ports:
36      - "22"
37
38 networks:
39   mpc_net:
40     driver: bridge
41     ipam:
42       driver: default
43       config:
44         - subnet: 192.168.1.0/24
45           gateway: 192.168.1.1
46

```

```

1 # Dockerfile based on: https://github.com/owidner/docker.openmpi/blob/master/Dockerfile
2 # https://codeburst.io/direct-connection-to-a-docker-container-with-ssh-56e1a02744ees
3 # Build this image: docker build -t mpc .
4
5 FROM ubuntu:20.04
6 MAINTAINER Pierre-Francois Wolfe <pwolfe@bu.edu>
7
8 ENV USER mpc
9 ENV HOME=/home/${USER}
10 ARG DEBIAN_FRONTEND=noninteractive
11
12 RUN apt update -y && \
13     apt-get install -y --no-install-recommends sudo apt-utils && \
14     apt-get install -y --no-install-recommends openssh-server \
15     make ssh gcc libopenmpi-dev openmpi-bin libsodium23 libsodium-dev && \
16     apt clean && \
17     apt purge && \
18     rm -rf /var/lib/apt/lists/* /tmp/* /var/tmp/*
19
20 RUN mkdir /var/run/ssh
21 RUN echo 'root:${USER}' | chpasswd
22 RUN sed -i 's/PermitRootLogin without-password/PermitRootLogin yes/' /etc/ssh/sshd_config
23
24 # SSH Login fix. Otherwise user is kicked off after login
25 RUN sed 's@session\s*required\s*pam_loginuid.so@session optional pam_loginuid.so@g' -i /etc/pam.d/ssh
26
27 ENV NOTVISIBLE "in users profile"
28 RUN echo "export VISIBLE=now" >> /etc/profile
29
30 # -----
31 # Add on 'mpc' user
32 # -----
33
34 RUN adduser --disabled-password --gecos "" ${USER} && \
35     echo "${USER} ALL=(ALL) NOPASSWD:ALL" >> /etc/sudoers
36

```

```

37 # -----
38 # Set-Up OpenSSH Key
39 # -----
40
41 ENV SSHDIR ${HOME}/.ssh/
42
43 RUN mkdir -p ${SSHDIR}
44
45 ADD ssh/config ${SSHDIR}/config
46 ADD ssh/id_rsa.mpi ${SSHDIR}/id_rsa
47 ADD ssh/id_rsa.mpi.pub ${SSHDIR}/id_rsa.pub
48 ADD ssh/id_rsa.mpi.pub ${SSHDIR}/authorized_keys
49
50 RUN chmod -R 600 ${SSHDIR}* && \
51     chown -R ${USER}:${USER} ${SSHDIR}
52
53 # -----
54 # Copy NPC code
55 # -----
56
57 COPY src/* /home/${USER}/code/src/
58 COPY experiments/* /home/${USER}/code/experiments/
59 COPY tests/* /home/${USER}/code/tests/
60 COPY launch.sh /home/${USER}/code/experiments/
61
62 WORKDIR /home/${USER}/code/experiments
63 RUN make exp-exchange
64
65 ENV OMPI_ALLOW_RUN_AS_ROOT=1
66 ENV OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1
67
68 RUN chown -R ${USER}:${USER} ${HOME}/code
69
70 RUN /usr/bin/ssh-keygen -A
71
72 EXPOSE 22
73
74 CMD ["/usr/sbin/sshd", "-D"]
75

```


Docker-compose to OpenShift

- Current (local)
 - docker-compose
 - Docker Containers
- In Progress
 - OpenShift (MOC)
 - Kubernetes
 - Docker Containers
- Conversion Tool
 - Kubernetes Kompose

docker-compose.yml

kompose --provider openshift --file
docker-compose.yml convert

party-0-service.yaml
party-1-service.yaml
party-2-service.yaml
party-0-deploymentconfig.yaml
party-0-imagestream.yaml
party-1-deploymentconfig.yaml
party-1-imagestream.yaml
party-2-deploymentconfig.yaml
party-2-imagestream.yaml

oc new-app scripts_party-0
scripts_party-1 scripts_party-2

APPLICATION scriptsparty-0				
>	DEPLOYMENT CONFIG scriptsparty-0	1 Error	No deployments for scriptsparty-0	:
>	DEPLOYMENT CONFIG scriptsparty-1	1 Error	No deployments for scriptsparty-1	:
>	DEPLOYMENT CONFIG scriptsparty-2	1 Error	No deployments for scriptsparty-2	:

Boston University CS & ECE

Launching on OpenShift WIP

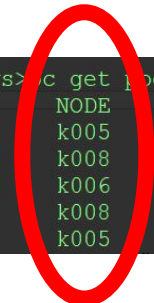
BOSTON
UNIVERSITY

OpenShift - Deploying Pods on Specific Nodes

When deploying multiple pods of same application, pods get assigned to different worker nodes randomly.

```
B:\Study\Boston University\Fall 2020\EC 528 - Cloud Computing\Project\openshift-client-windows>oc get pods -o wide
```

NAME	READY	STATUS	RESTARTS	AGE	IP	NODE
docker101tutorial-794f8f8dd8-c5mlb	1/1	Running	1	34d	10.128.1.34	k005
docker101tutorial-794f8f8dd8-dlgm8	1/1	Running	1	34d	10.128.8.101	k008
docker101tutorial-794f8f8dd8-js4ww	1/1	Running	1	34d	10.128.4.189	k006
docker101tutorial-794f8f8dd8-wljmd	1/1	Running	1	34d	10.128.8.93	k008
docker101tutorial-794f8f8dd8-zkzkh	1/1	Running	1	34d	10.128.1.33	k005



Running all MPC parties pods on the same nodes may have different latencies/computation delays as is the case otherwise.

OpenShift - Deploying Pods on Specific Nodes

Tried specifying what nodes to place pods on.

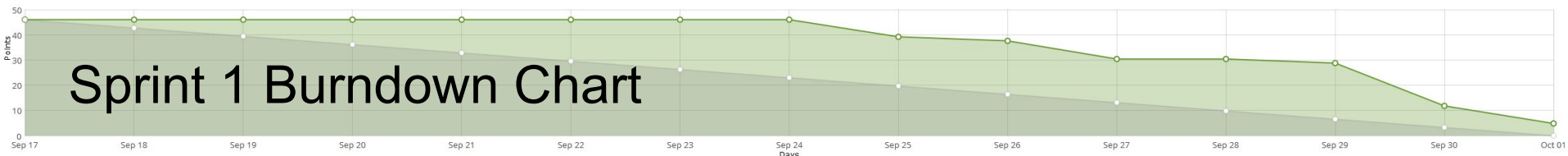
Failed → Most documentation for this is for OpenShift 3, while the latest version we're using is OpenShift 4.

An authorization error is encountered: Admin access / elevated rights needed !

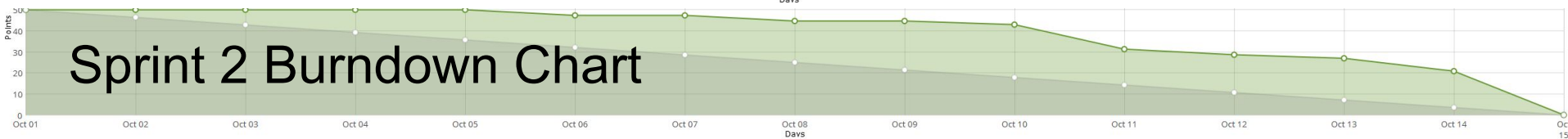
```
B:\Study\Boston University\Fall 2020\EC 528 - Cloud Computing\Project\openshift-client-windows>oc get nodes
Error from server (Forbidden): nodes is forbidden: User "hasnain@bu.edu" cannot list nodes at the cluster scope: no RBAC policy matched

B:\Study\Boston University\Fall 2020\EC 528 - Cloud Computing\Project\openshift-client-windows>oc edit namespace
Error from server (Forbidden): namespaces is forbidden: User "hasnain@bu.edu" cannot list namespaces at the cluster scope: no RBAC policy matched
```

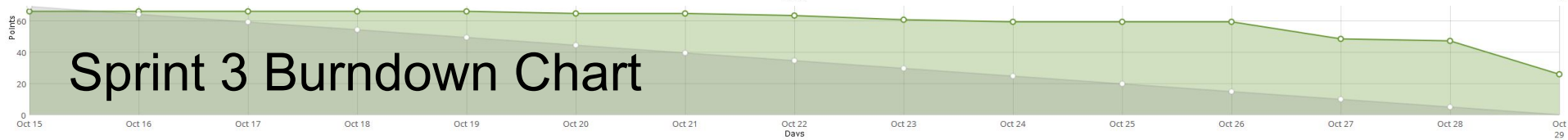
Sprint 1 Burndown Chart



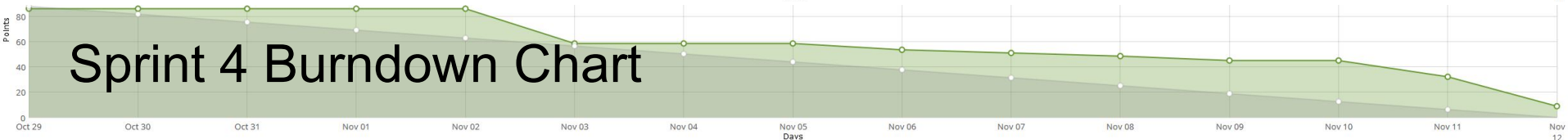
Sprint 2 Burndown Chart



Sprint 3 Burndown Chart



Sprint 4 Burndown Chart



Sprint 5 - Some Known Stories

- As a researcher, I want to use a comprehensive test suite across all deployments...
 - Exp-exchange: iterate over message size, transaction size, other MPI options...
 - Add score-p wrapper toggle (and use some MPI tools for insights)
 - Push all changes to test setups (CloudLab *.tar.gz, rebuild Docker, rsync to VMs)
- As a researcher, I want to deploy final tests to each environment of interest.
 - Docker containers on OpenShift (finish debugging)
 - Revisit existing VM setup on OpenStack
 - Ring topology on CloudLab (if possible, otherwise LAN)
- As a researcher, I want to compare the same tests run on different platforms.
 - Conclusions about best performance of tested methods
 - Documentation and any other insights

Thank you

...any questions?