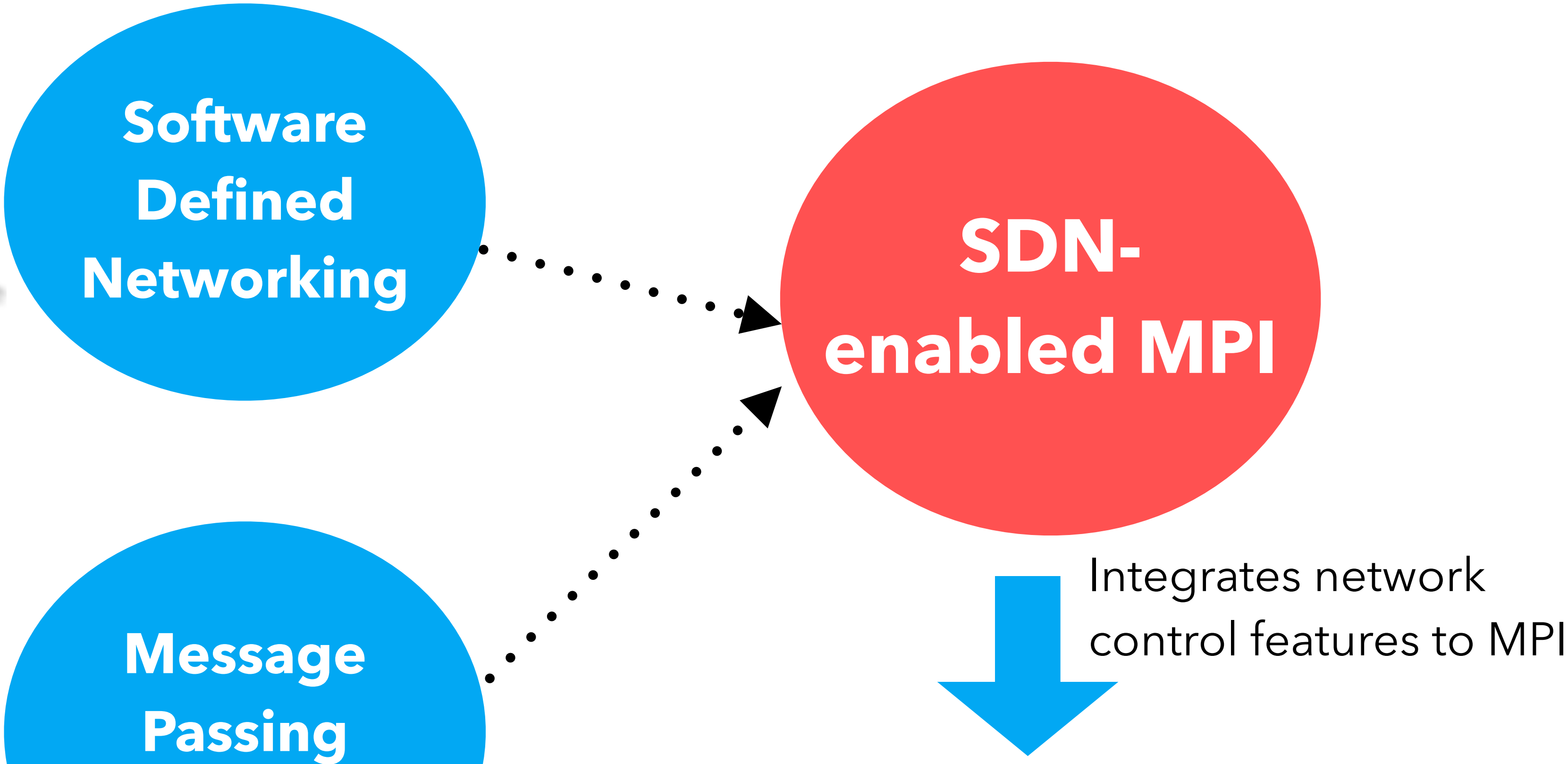
Control Sequence Generator for Generic SDN-enabled MPI Framework

Keichi Takahashi, Baatarsuren Munkhdorj, Khureltulga Dashdavaa,
Susumu Date, Yoshiyuki Kido, Shinji Shimojo
Cybermedia Center, Osaka University

Towards SDN-enabled MPI

OPEN NETWORKING FOUNDATION



Software Defined Networking

Message Passing Interface

SDN-enabled MPI

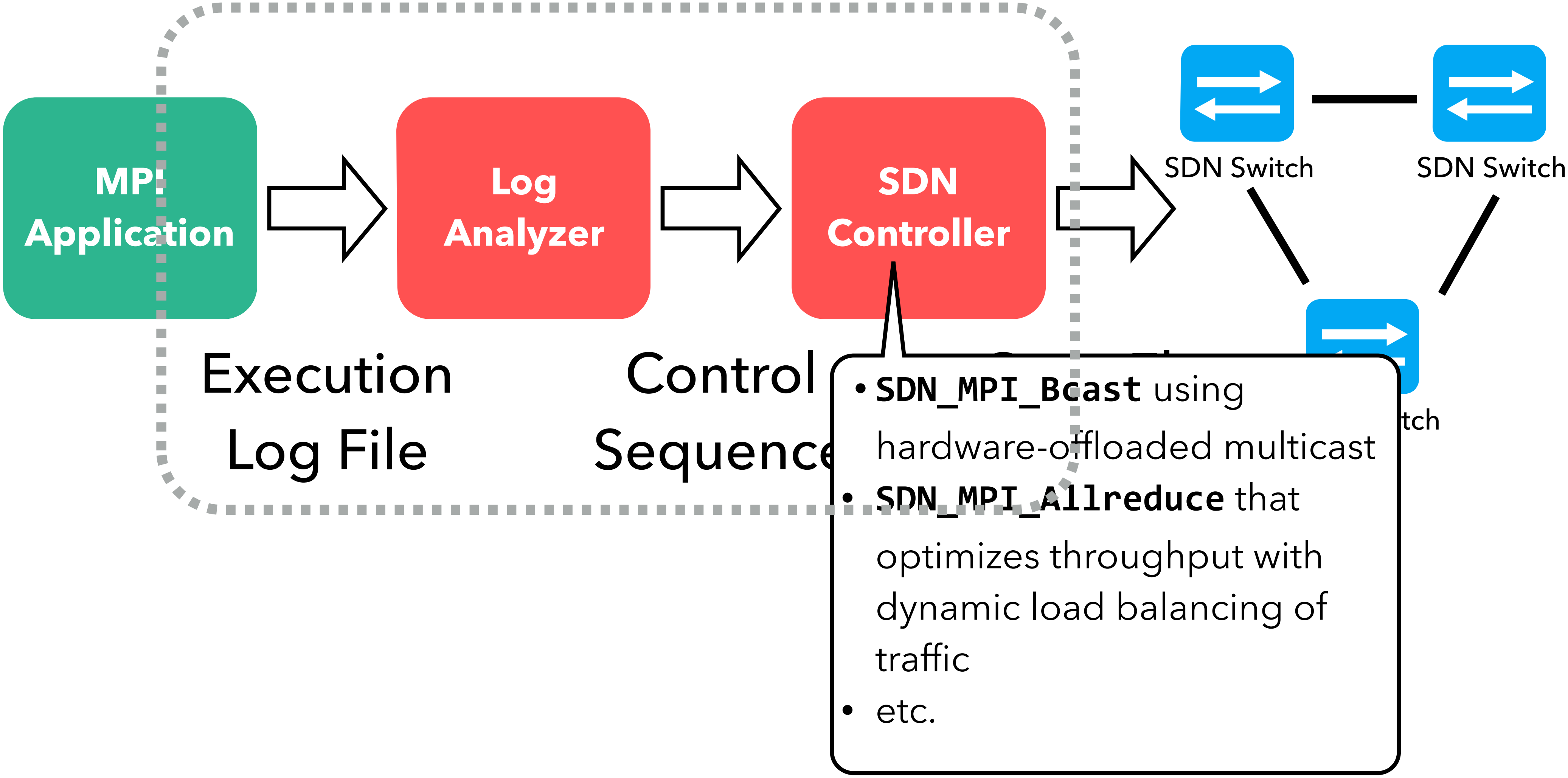
Integrates network control features to MPI

OPEN MPI

MVAPICH

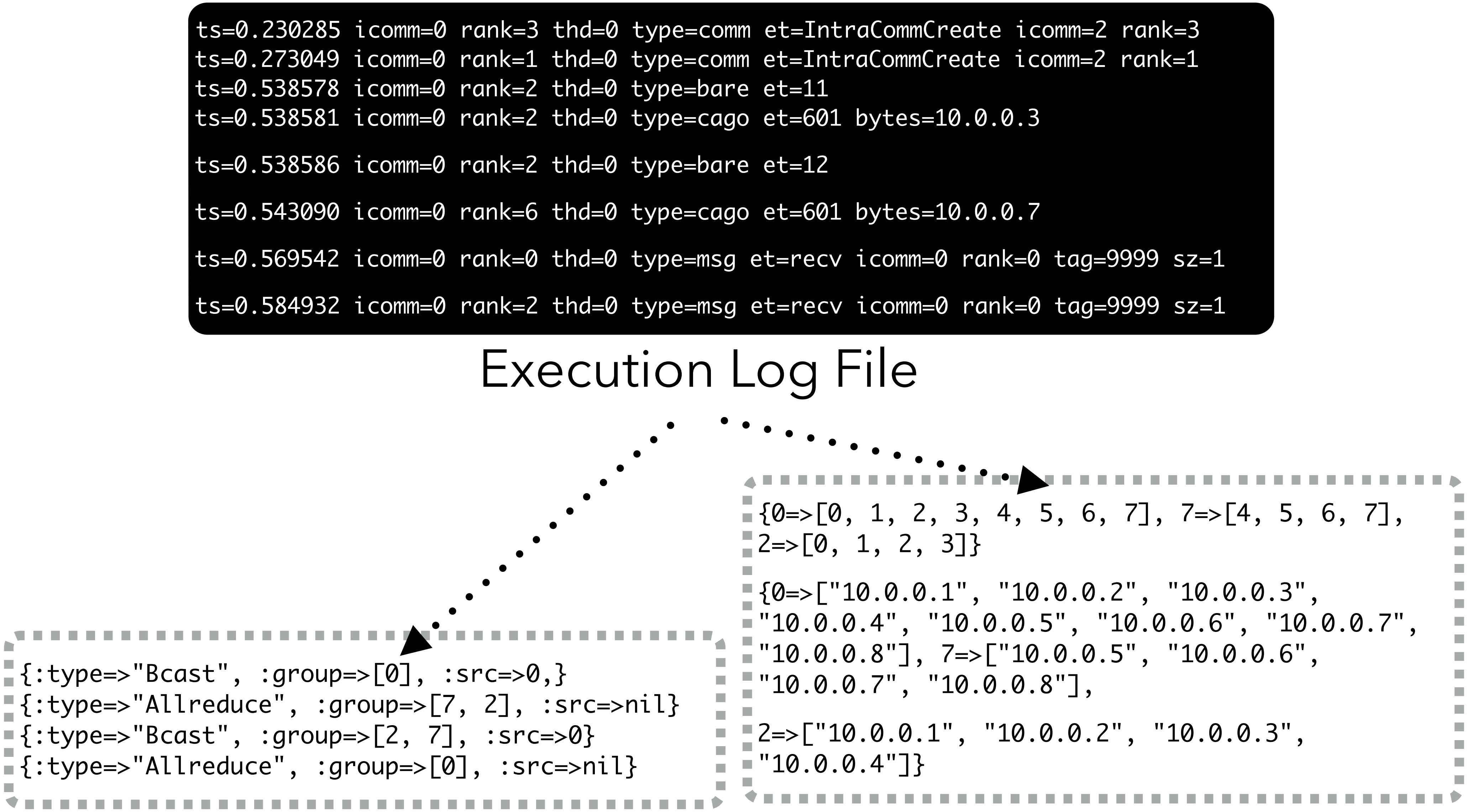
- Hardware multicast
- Dynamic network load balancing
- Link congestion avoidance
- etc.

Generic SDN-enabled MPI Framework



- **`SDN_MPI_Bcast`** using hardware-offloaded multicast
- **`SDN_MPI_Allreduce`** that optimizes throughput with dynamic load balancing of traffic
- etc.

Control Sequence Generator



```
ts=0.230285 icomm=0 rank=3 thd=0 type=comm et=IntraCommCreate icomm=2 rank=3
ts=0.273049 icomm=0 rank=1 thd=0 type=comm et=IntraCommCreate icomm=2 rank=1
ts=0.538578 icomm=0 rank=2 thd=0 type=bare et=11
ts=0.538581 icomm=0 rank=2 thd=0 type=cago et=601 bytes=10.0.0.3

ts=0.538586 icomm=0 rank=2 thd=0 type=bare et=12

ts=0.543090 icomm=0 rank=6 thd=0 type=cago et=601 bytes=10.0.0.7

ts=0.569542 icomm=0 rank=0 thd=0 type=msg et=recv icomm=0 rank=0 tag=9999 sz=1

ts=0.584932 icomm=0 rank=2 thd=0 type=msg et=recv icomm=0 rank=0 tag=9999 sz=1
```

Execution Log File

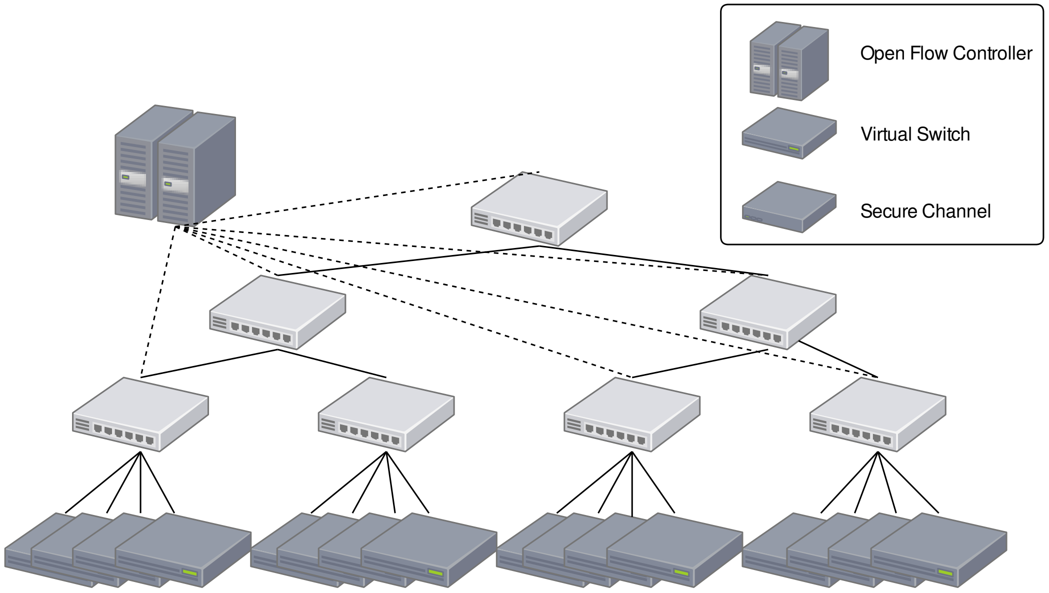
```
{:type=>"Bcast", :group=>[0], :src=>0,}
{:type=>"Allreduce", :group=>[7, 2], :src=>nil}
{:type=>"Bcast", :group=>[2, 7], :src=>0}
{:type=>"Allreduce", :group=>[0], :src=>nil}
```

Event List

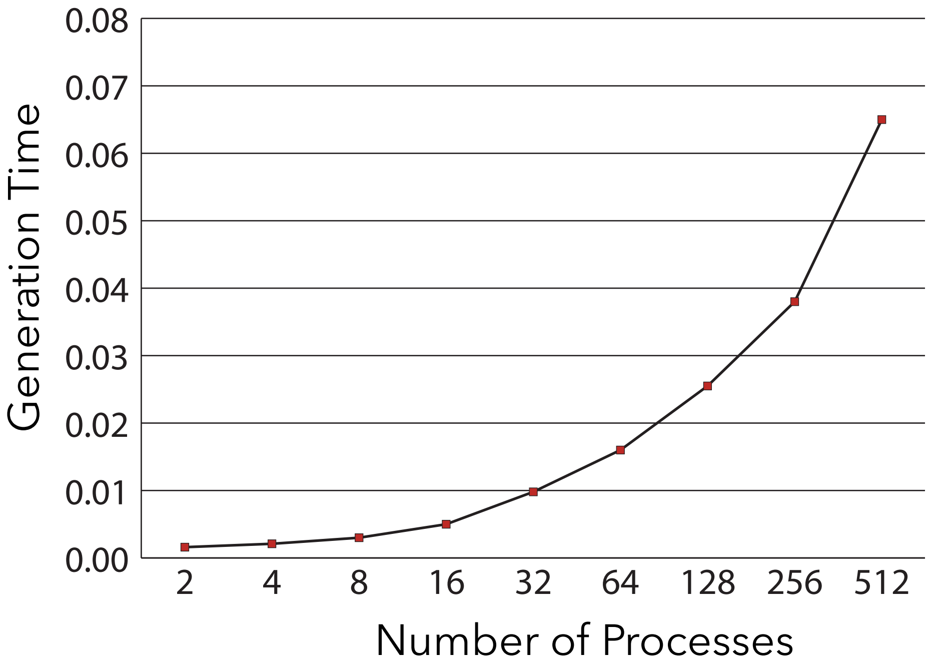
```
{0=>[0, 1, 2, 3, 4, 5, 6, 7], 7=>[4, 5, 6, 7],
2=>[0, 1, 2, 3]}

{0=>["10.0.0.1", "10.0.0.2", "10.0.0.3",
"10.0.0.4", "10.0.0.5", "10.0.0.6", "10.0.0.7",
"10.0.0.8"], 7=>["10.0.0.5", "10.0.0.6",
"10.0.0.7", "10.0.0.8"],

2=>["10.0.0.1", "10.0.0.2", "10.0.0.3",
"10.0.0.4"]}
```

Process/Group Attributes

Prototype Implementation and Initial Experimental Results



Virtual Cluster
Emulated with **Mininet**



Control Sequence Generation

Distinct SDN-enabled MPI functions/modules can be used simultaneously