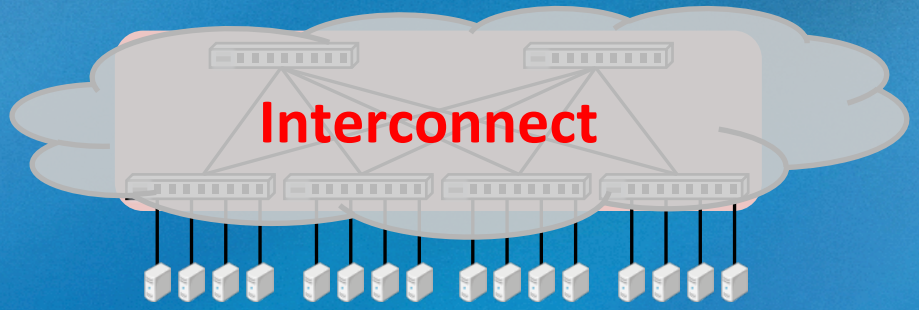
Toward Flexible and Efficient Computing Resource Service by SDN-enhanced Job Management System Framework

Yasuhiro Watashiba [†1], Susumu Date [†2], Hirotake Abe [†3], Yoshiyuki Kido [†2], Kohei Ichikawa [†1], Hiroaki Yamanaka [†4], Eiji Kawai [†4], and Shinji Shimojo [†2,4]

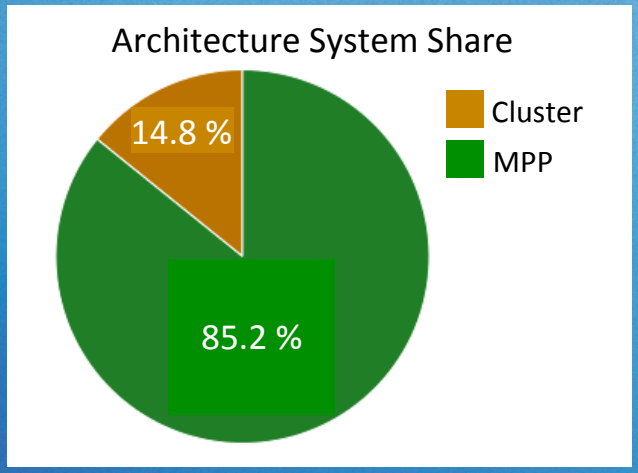
†1 Nara Institute of Science and Technology, Japan
†2 Osaka University, Japan
†3 University of Tsukuba, Japan
†4 National Institute of Information and Communications Technology (NICT), Japan

High-performance Computing Environment

- Current high-performance computing environment
  - The dominant trend is cluster system.



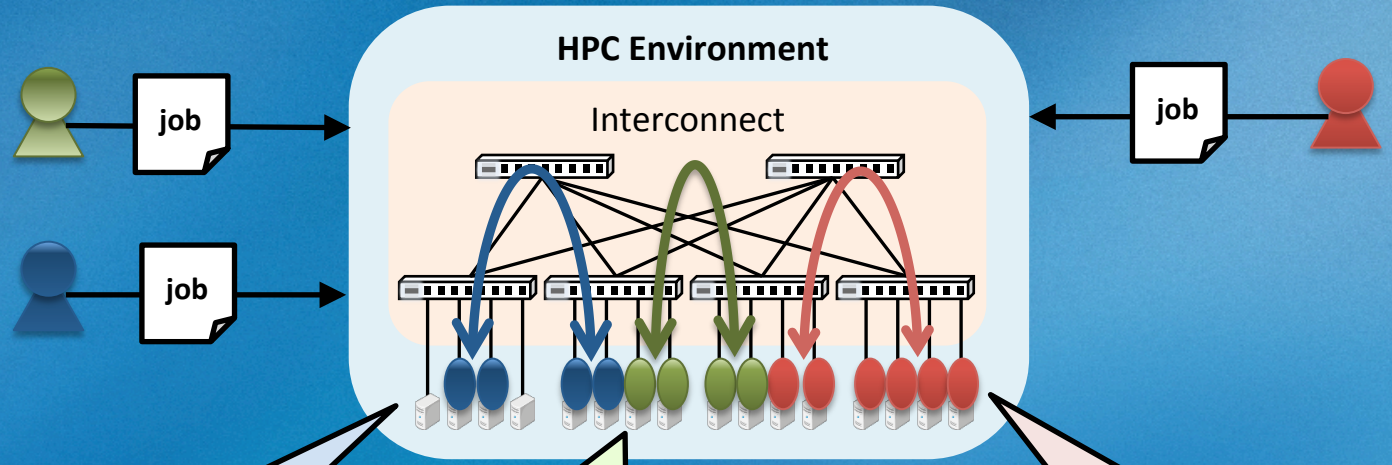
- Performance improvement
  - Increasing computing nodes
    => Scaling up interconnect
  - Equipping accelerator



Statistics of TOP500 Supercomputer Sites
(November 2015)

Resources have become diversified and complex.

Requirement for Resource Management



Efficient execution of parallel and distributed computation for gaining high computational performance.

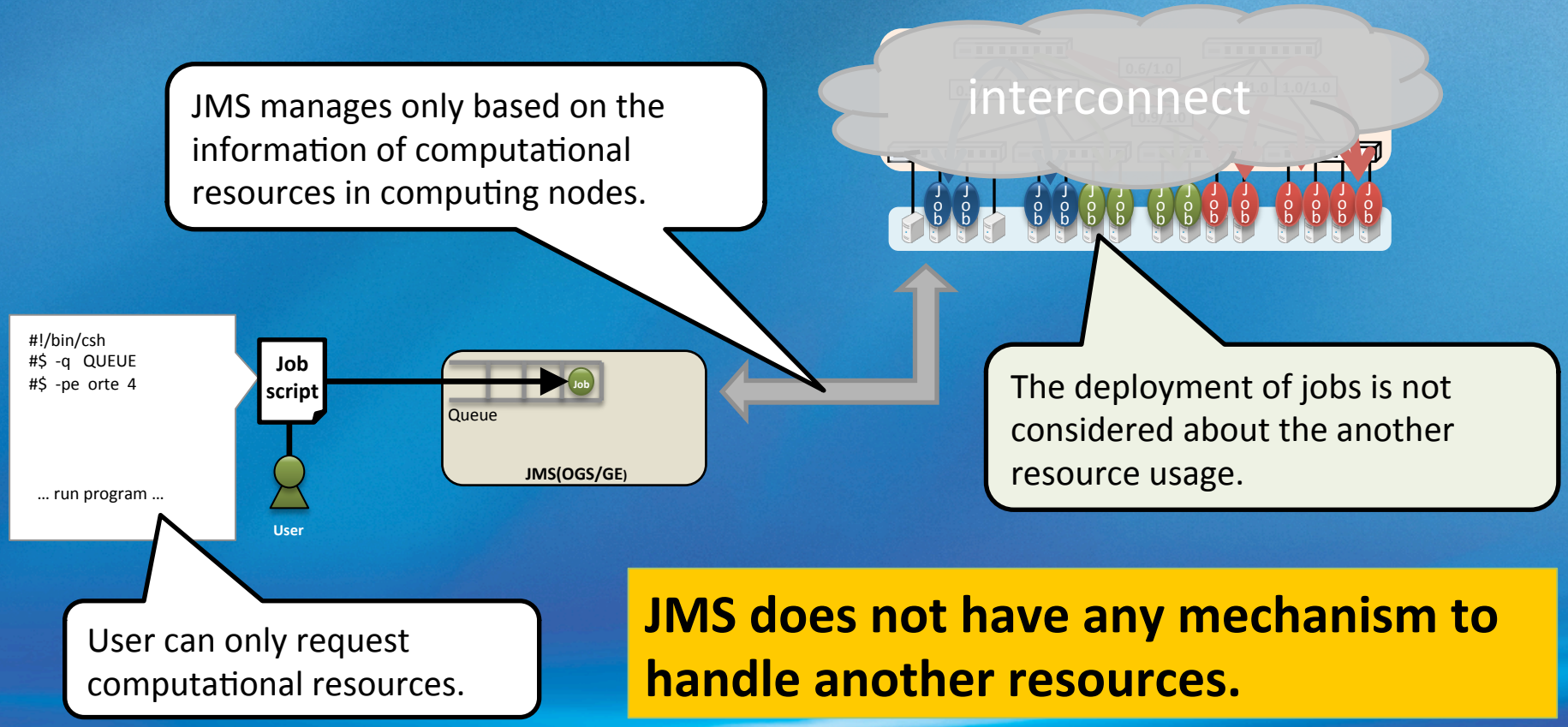
Concurrently running as many computations as possible for achieving high job throughput.

Allocating an appropriate set of resources for handling various computational request.

Efficient and flexible resource management system is necessary for achieving these requirements.

Traditional Resource Management

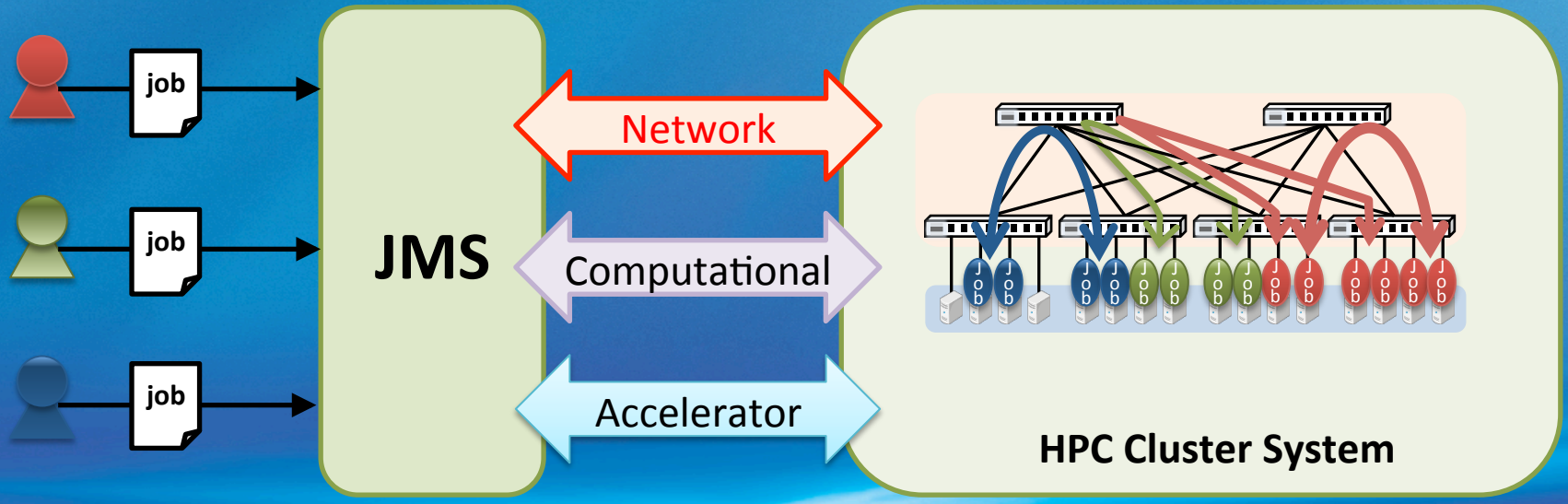
- Job Management System (JMS)
  - Traditional Resource Management System on HPC cluster systems
  - Managing and allocating only computational resources.



JMS does not have any mechanism to handle another resources.

Challenges

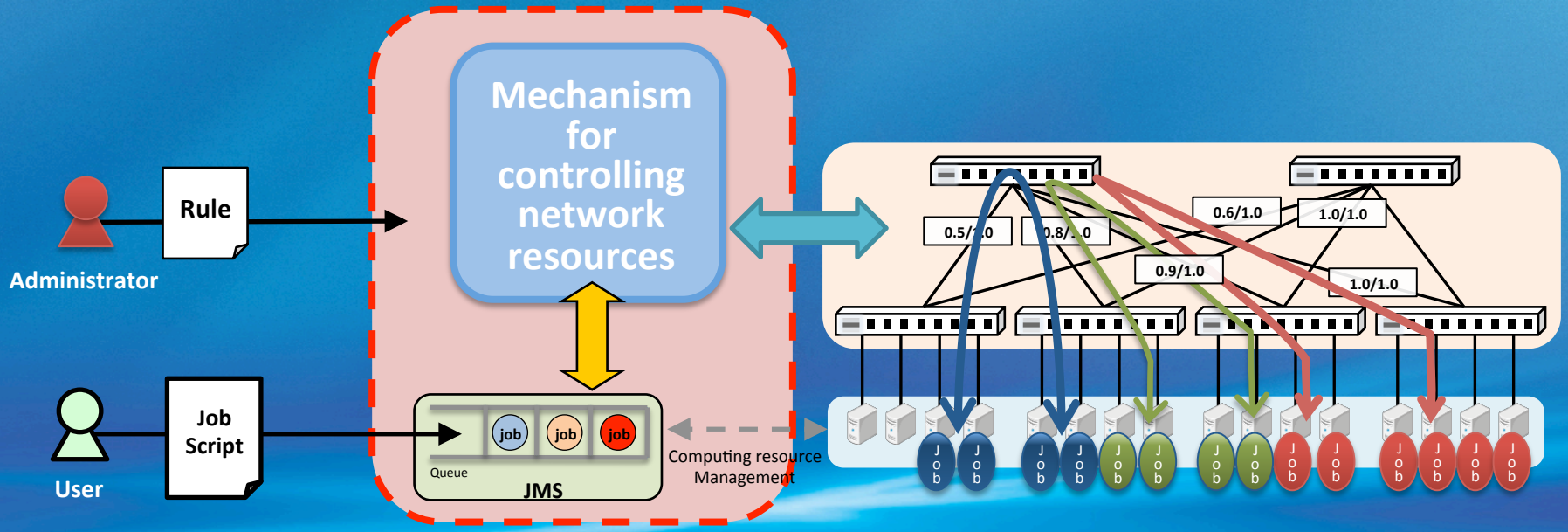
Realizing mechanisms for handling various resources on JMS.



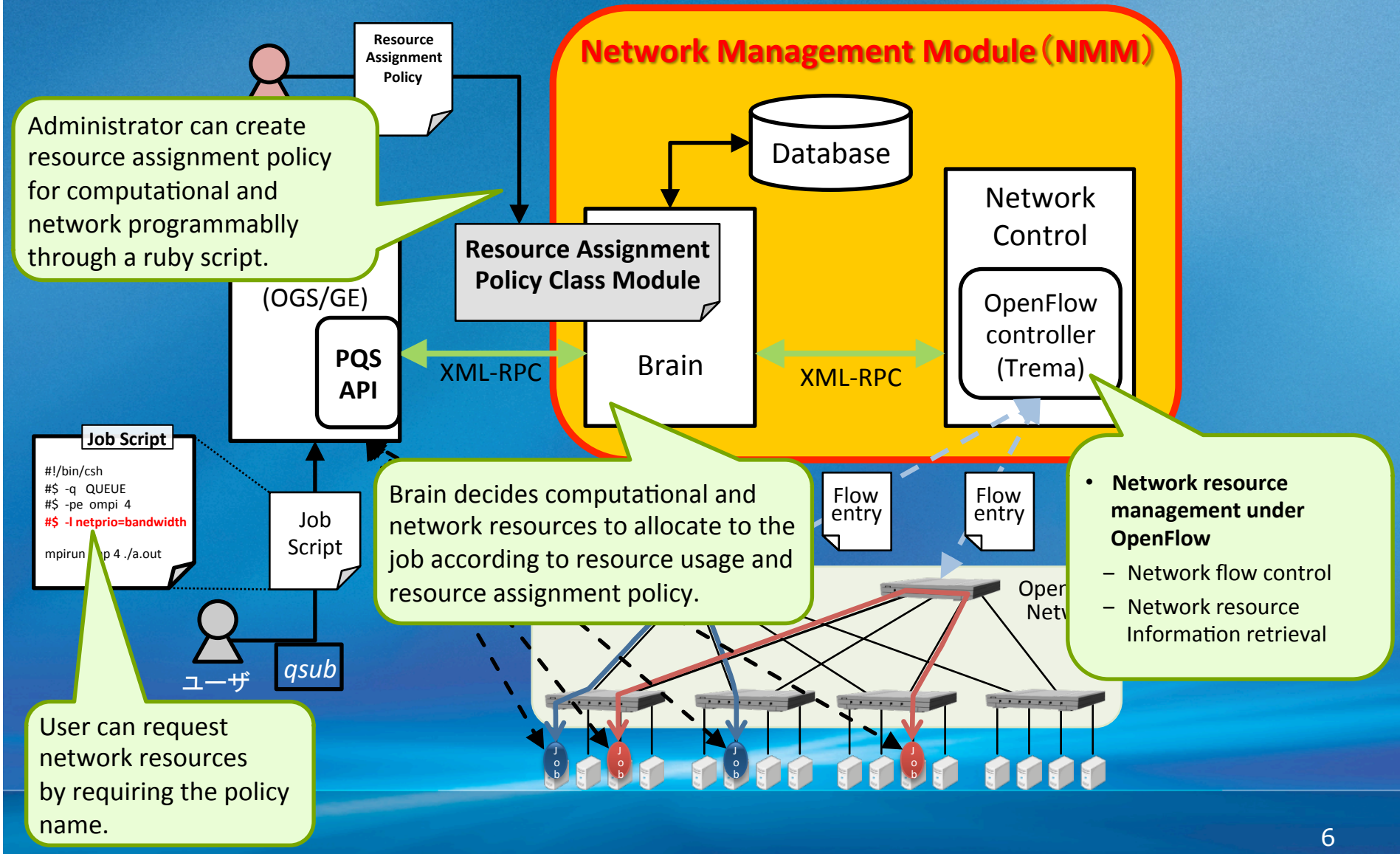
Design of Network-aware JMS

➢Functional Requirements

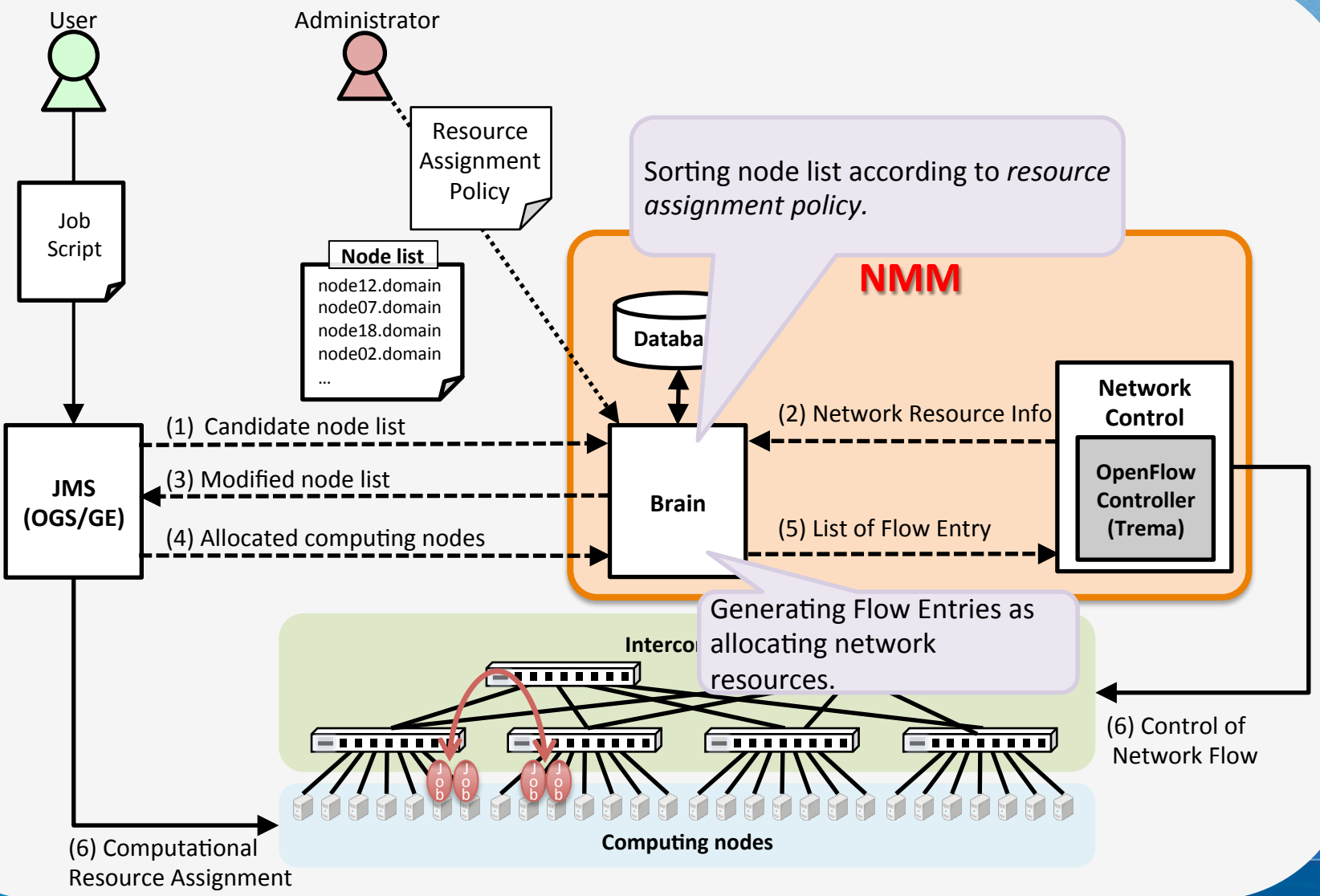
1. Understanding the information of network resources.
2. Allocating network resources explicitly.
3. User interface for resource request.
4. Deciding resource allocation by taking into consideration both computational and network resources.
5. Administrator interface for making a resource allocation rule.



SDN-enhanced JMS Framework (Ver.1)



Flow of SDN-enhanced JMS Framework



Example of job script for SDN-enhanced JMS Framework

```
#!/bin/csh
#$  -q   QUEUE
#$  -pe  ompi  4
#$  -l netprio=policy_name

mpirun  -np $NSLOTS  ./a.out
```

Example of Resource Assignment Policy

```
## Resource Assignment Policy

sort_qlist {
   Get a pick list of computing nodes.
   Get network information.
   if netprio is "policy_name"
      Calculate appropriate set of resources
      Sort the list according to the result
   else if …
      …
   end
   Return the pick list to OGS/GE
}

set_route {
   Get a computing node list.
   Get network information.
   Generate flow entries between computing nodes in the list.
   Return the flow entries to Network Control module.
}

del_route {
   Get flow entries assigned to a target job
   Remove the flow entries
}
```

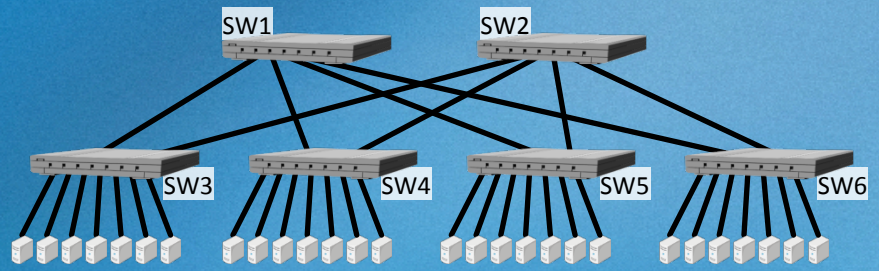
Demo

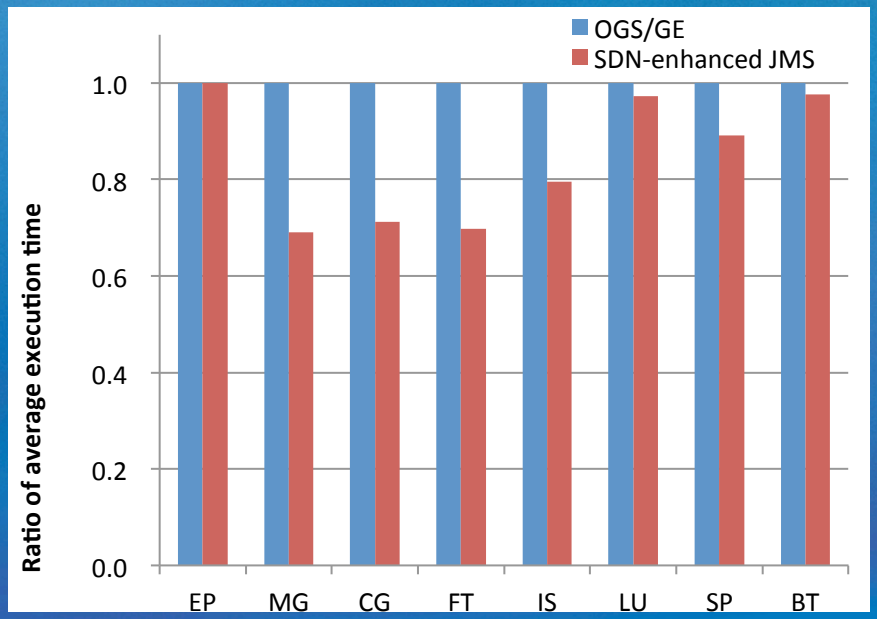
1. Submitting sample job
2. Monitor command for network resource allocation

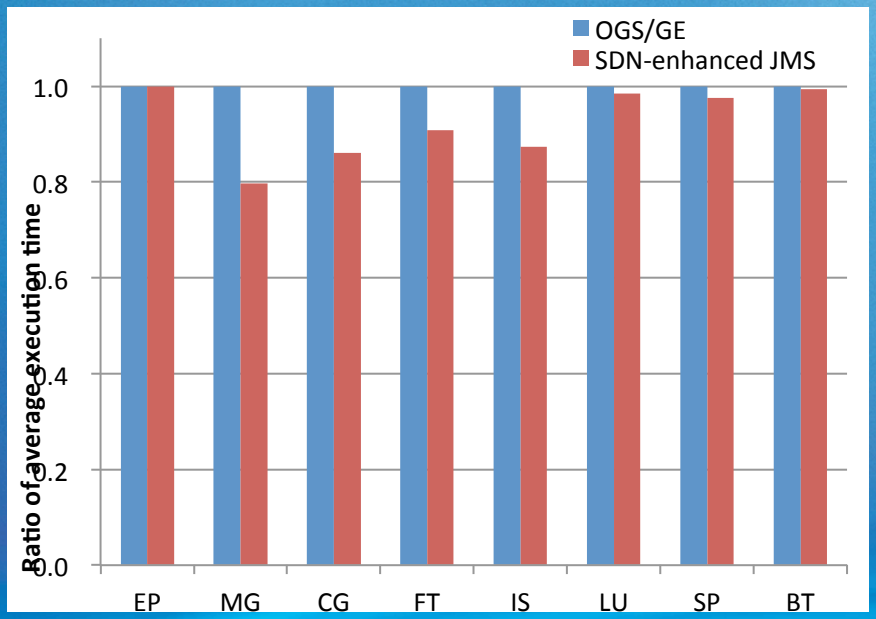
```
Every 2.0s: qstat -f; echo ''; qrstat; qstat+ -n -c                    Mon Sep 22 14:23:05 2014

queuename                      qtype resv/used/tot. load_avg arch          states
---------------------------------------------------------------------------------
all.q@pepsi13.default.domain   BIP   0/1/24         0.00     linux-x64
    735 0.55500 mpitest.sh sge          r     09/22/2014 14:22:28     1
---------------------------------------------------------------------------------
all.q@pepsi14.default.domain   BIP   0/1/24         0.08     linux-x64
    735 0.55500 mpitest.sh sge          r     09/22/2014 14:22:28     1
---------------------------------------------------------------------------------
all.q@pepsi27.default.domain   BIP   0/0/24         0.00     linux-x64
---------------------------------------------------------------------------------
all.q@pepsi28.default.domain   BIP   0/0/24         0.00     linux-x64

connection          job_id                          bandwidth
-------------------------------------------------------------------------------------------
pepsi14.default.domain -> ofs101 735                          75332.4 / 1000000000.0 (bps)
-------------------------------------------------------------------------------------------
ofs101 -> pepsi14.default.domain 735                          3691277.9 / 1000000000.0 (bps)
-------------------------------------------------------------------------------------------
pepsi13.default.domain -> ofs101 735                          1174.2 / 1000000000.0 (bps)
-------------------------------------------------------------------------------------------
ofs101 -> pepsi13.default.domain 735                          5086.4 / 1000000000.0 (bps)
```

Evaluation results

- NAS Parallel Benchmarks
  - 2 kind of cluster system with same interconnect topology and different bandwidth.



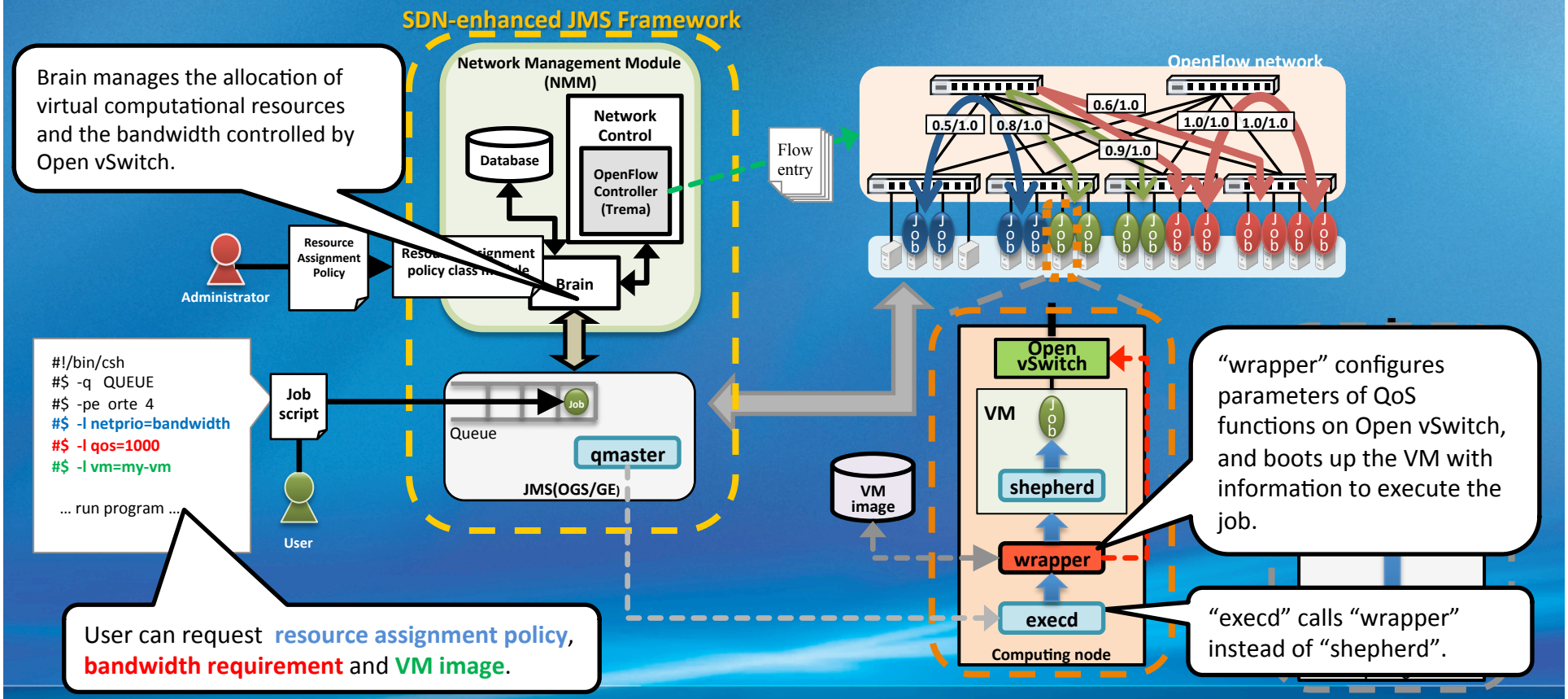




Left graph: MG 30.9%, CG 28.8T, FT 30.2%, IS 20.4%
Right graph: MG 20.2%, CG 13.9%, FT 9.2%, IS 12.6%

Virtualized Computational Resource Management

- **SDN-enhance JMS Framework Ver.2**
  - **Boot up user-request VM and deploy process of job on the VM**
  - **Configure parameters of QoS functions on Open vSwitch**



Toward SDN + HPC Environment

Paradigm Shift from traditional HPC pursuing computing performance under the "static" uncontrollable network resource to new HPC pursuing computing performance taking advantage of "dynamic" controllable network resource.

(1) Resource Management that treats interconnect as controllable resource as well as computing resources such as CPU and memory.

(4) On-demand cloud formation aggregating computational resources from multiple sites

Computing request

Virtual cluster

proc ess

proc ess

proc ess

proc ess

Interconnect controller

In-site SDN

(2) Run-time dynamic control of Inter-process communication

Visualization Request

Inter-site SDN

Inter-site SDN controller

(3) Flow control based on network requirements changes by user interactive operations for visualization

Toward SDN + HPC Environment

…ting performance under the "static"
…ing computing performance taking
…rce.

(1) Resources of Osaka University are planed to provide via the SDN-enhanced JMS Framework Ver.1.

(3) Communication and collaboration for resource management technology.

(2) Source code of the SDN-enhanced JMS Framework will be available from our homepage.

- WEB page:
  http://hpc-sdn.ime.cmc.osaka-u.ac.jp/sdnjms/

- Contact:
  watashiba@is.naist.jp