Cybermedia Center
Osaka University

Cybermedia Center

Osaka University, Japan
SC22 BOOTH 1613

ns-3-based Interconnect Simulator for Network Simulation with Job Scheduling

Background : Aim of Interconnect Design in Supercomputing Systems is Changing

A variety of jobs are performed on today's supercomputing systems. The number of compute nodes requested by such jobs is diverse and then much inter-node communication take place.
⇨ Interconnects of supercomputing systems should be **designed using simulators to examine the performance in communication.**

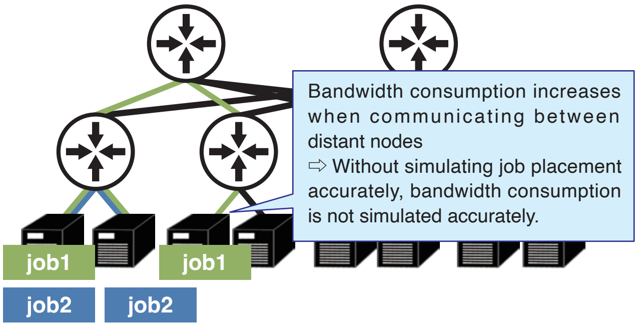
Traditional Supercomputing Systems	
Expected Workload	Computation-intensive MPI (Message Passing Interface) jobs.
Aim of Interconnect Design	Focus on the cost to increase the number of compute nodes.
Method of Interconnect Design	･Select from stable and mature technologies such as Fat-tree and ECMP. ･Parameters are determined by calculations and other simple estimates.

Next Supercomputing System	
Expected Workload	Communication-intensive jobs using distributed processing frameworks.
Aim of Interconnect Design	Focus on the performance to accelerate inter-node communication.
Method of Interconnect Design	･Select from stable and mature technologies and/or state-of-the-art technologies such as DragonFly and adaptive routing. ･Parameters are determined by simulations to examine interconnect performance.

Problem: The Effects of Job Scheduling Are Missed by Existing Network Simulators

When simulating interconnects in a supercomputing system, the simulation result is incorrect in the case of using only existing network simulators. The reason is **existing network simulators cannot reproduce job placement by job schedulers.**

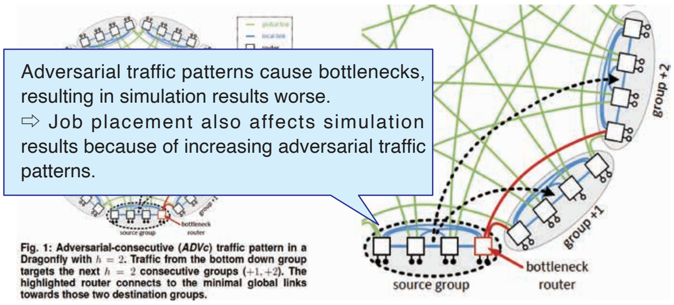
- **Traffic patterns* are changed by job placement.**
(*A set of communications within a certain time period)



Bandwidth consumption increases when communicating between distant nodes
⇨ Without simulating job placement accurately, bandwidth consumption is not simulated accurately.

Example of link capacity consumption depending on job placement

- **Adversarial traffic patterns**** cause misunderstanding of the network performance. (**Traffic patterns that degrade network performance)



Adversarial traffic patterns cause bottlenecks, resulting in simulation results worse.
⇨ Job placement also affects simulation results because of increasing adversarial traffic patterns.

Fig. 1: Adversarial-consecutive (*ADVc*) traffic pattern in a Dragonfly with $h = 2$. Traffic from the bottom down group targets the next $h = 2$ consecutive groups $(+1, +2)$. The highlighted router connects to the minimal global links towards those two destination groups.

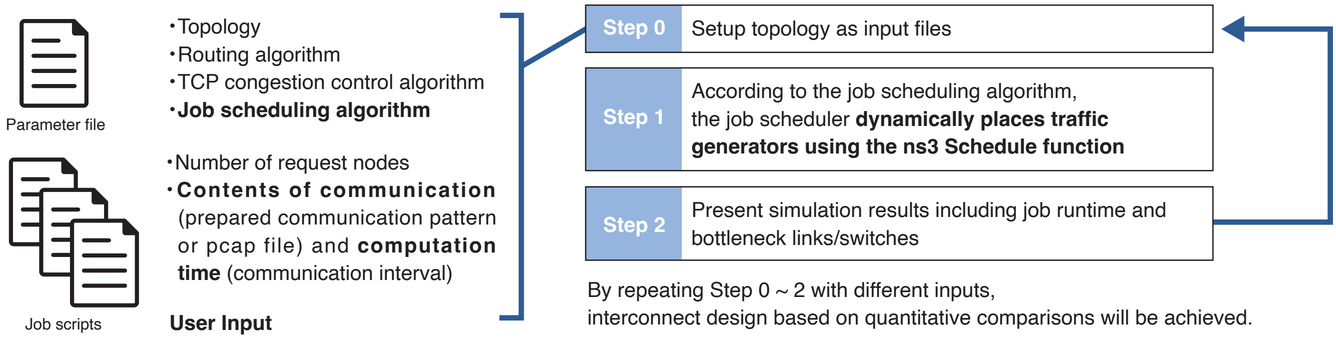
Adversarial traffic patterns in DragonFly topology [1]

[1] P. Fuentes, E. Vallejo, C. Camarero, R. Beivide and M. Valero, "Throughput Unfairness in Dragonfly Networks under Realistic Traffic Patterns," 2015 IEEE International Conference on Cluster Computing, 2015, pp. 801-808, doi: 10.1109/CLUSTER.2015.136.

Proposal : ns-3-based Interconnect Simulator for Interconnect Design (In-Progress)

To achieve network simulation with job scheduling, we decided to implement a job scheduling function as a module for ns-3.

- **Assets**: Interconnect research results are implemented in ns-3, **enabling simulations using state-of-the-art technologies.**
- **Expandability**: ns-3 is modularized, making it **easy to expand the job scheduling** functionalities.
- **Packet-level simulation**: accurate simulation of network latency should **reduce performance estimation errors.**



- Topology
- Routing algorithm
- TCP congestion control algorithm
- **Job scheduling algorithm**

- Number of request nodes
- **Contents of communication** (prepared communication pattern or pcap file) and **computation time** (communication interval)

User Input

Step 0 Setup topology as input files

Step 1 According to the job scheduling algorithm, the job scheduler **dynamically places traffic generators using the ns3 Schedule function**

Step 2 Present simulation results including job runtime and bottleneck links/switches

By repeating Step 0 ~ 2 with different inputs,
interconnect design based on quantitative comparisons will be achieved.

Contact : sc22@ais.cmc.osaka-u.ac.jp https://www.cmc.osaka-u.ac.jp/