Network-aware Job Scheduling for the HPC Cloud

2012/12/11

Yoshitomo Murata
Cyberscience Center, Tohoku University
Agenda

- Background
- Vector Computing Cloud
- Network-aware Job Scheduling
- Evaluation
- Conclusions
High Performance Computing (HPC) requires more computing resources.
- HPC is used in many fields such as research and development.

Distributed computing systems provide high-computing power to HPC users
- Multi computing resources coordinate on resolving large-scale problems.
- Ex) Grid Computing, Volunteer Computing
Distributed HPC Systems

16th Workshop on Sustained Simulation Performance (WSSP)
Vector Computing Cloud
Advantages of Vector Computing Cloud

- **Load-Balancing**
  - The submitted job is automatically assigned to low-load node.
  - High utilization of multiple systems is realized.

- **Realizing Virtual Parallel Computer**
  - The virtual parallel computer is not limited hardware configuration of each site.

Improving the flexibility of job execution and resource management
Composition of Wide-Area Vector co-operation Environment

- Two supercomputer sites
  - Two SX-9 nodes at Tohoku University
  - Two SX-9 nodes at Osaka University

- Two high-speed wide area network
  - SINET-3 (1Gbps)
  - JGN2Plus (10Gbps)
Potential of Wide-Area Vector co-operation Environment - HPL Benchmark

Inhouse Performance
Wide area co-operation Performance

16th Workshop on Sustained Simulation Performance (WSSP)

Efficiency 51.3%
1.6TFlop/s
single node
Peak performance
Potential of Wide-Area Vector co-operation Environment - Real Application

The fluid flow around a cylinder is simulated by a particle method on the wide-area environment.

<table>
<thead>
<tr>
<th>Prob. size</th>
<th># of node</th>
<th>Wide-area</th>
<th>local</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 x 16</td>
<td>1 x 1</td>
<td>3,886</td>
<td>613</td>
</tr>
<tr>
<td></td>
<td>2 x 2</td>
<td>3,777</td>
<td>392</td>
</tr>
<tr>
<td>32 x 32</td>
<td>1 x 1</td>
<td>20,096</td>
<td>7,675</td>
</tr>
<tr>
<td></td>
<td>2 x 2</td>
<td>24,995</td>
<td>4,406</td>
</tr>
</tbody>
</table>

The visualized results shows the vortexes behind the cylinder. The wide-area vector co-operation has the ability for executing MPI programs successfully.
Job Scheduling on Vector Computing Cloud

- Scheduling Problem
  - There is no standard to select an appropriate site from distributed resources.
    
    **History-based job scheduling**

- The execution time of parallel application is limited by the performance of wide-area network.
  - To reduce the limitation by the wide-area network, new network selection mechanism is required.
History-based Job Scheduling

- Scheduling Problem
  - Difference in the site operation policies
    - Osaka University: Reservation-based
      - Guarantee of the job-start time
      - Time restriction of the job-execution
    - Tohoku University: Queuing-based
      - High usability of the computing resources
      - No guarantee of the job-start time
  - No fairness and efficiency in the job scheduling between sites.
History-based Job Scheduling

- Assigning a job to the site which has least waiting-time
- Achieving fairness and efficiency resource selection

Log of job-execution times

<table>
<thead>
<tr>
<th>command</th>
<th>Execution-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>100 sec</td>
</tr>
<tr>
<td>BBB</td>
<td>500 sec</td>
</tr>
<tr>
<td>CCC</td>
<td>200 sec</td>
</tr>
</tbody>
</table>

List of queued jobs

<table>
<thead>
<tr>
<th>order</th>
<th>command</th>
<th>wait-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AAA</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>CCC</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>BBB</td>
<td>300</td>
</tr>
<tr>
<td>3</td>
<td>AAA</td>
<td>800</td>
</tr>
</tbody>
</table>
Job Scheduling on Vector Computing Cloud

- Scheduling Problem
  - There is no standard to select an appropriate site from distributed resources.
  - **History-based job scheduling**

  - The execution time of parallel application is limited by the performance of wide-area network.
  - To reduce the limitation by the wide-area network, new network selection mechanism is required.
  - **Network-aware job scheduling**
NETWORK-AWARE JOB SCHEDULING
Network-aware Job scheduling

- **Objective**
  - Realizing the node and network allocation to minimize the influence of network in the wide-area co-operation environment.

- **Approach**
  - Modeling the execution time of parallel-job
    - Considering the network performance
  - Designing the node and network allocation mechanism
Modeling the Job Execution Time

- Presupposition
  - The changes of execution time depends on only network performance
    - Vector Computing Cloud consists of same kind of vector processors

- Equation
  \[ T_{exec} = T_{comp} + T_{comm} \]
  \[ T_{comp} \]: The time spent for computation (constant)
  \[ T_{comm} \]: The time spent for communication (variable)
Modeling the Network Communication Time

- Presuppositions
  - The communication time is decided on only the network bandwidth and latency.
    - the larger size of data is, the longer the communication time is.
    - the more count of communication is, the longer the communication time is.

- Equation

\[ T_{\text{comm}} = C_{\text{band}} F_{\text{band}} + C_{\text{lat}} F_{\text{lat}} \]

- \( C_{\text{band}} \): The normalized communication time regarding network bandwidth
- \( C_{\text{lat}} \): The normalized communication time regarding network latency
- \( F_{\text{band}} \): Factor of network bandwidth
- \( F_{\text{lat}} \): Factor of network latency (local network has 1.0, wide area network has more than 1.0)
Processes of Job Scheduling

Initial Step: Scheduler push the network information $F_{\text{band}}$ and $F_{\text{lat}}$ of each path to DB.

Start: User submits a job.

Step1: Scheduler retrieves the profiling information $T_{\text{comp}}$, $C_{\text{band}}$ and $C_{\text{lat}}$ from DB.

Step2: History-based job scheduler estimates the waiting time of each node.

Step3: Scheduler makes combinations of nodes which can start processes at the same time.

Step4: Scheduler retrieves the network information $F_{\text{band}}$ and $F_{\text{lat}}$ corresponding to above node combinations.

Step5: Scheduler calculates the make span of each combination, and selects the combination which has the shortest make span.

<table>
<thead>
<tr>
<th>Site</th>
<th>Waiting Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site A</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Site B</td>
<td>1 hour</td>
</tr>
<tr>
<td>Site C</td>
<td>1 hour</td>
</tr>
</tbody>
</table>

Network-aware job scheduler

History-based job scheduler

DB
Job Assignment Mechanism

Job submitting

Network

Resource Requirement

scheduler

Network Requirement

Application Profile

DB

OpenFlow

Network Routing
PERFORMANCE EVALUATION
## Simulation Parameters (1)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{comp}}$ ( frequency ) [sec.]</td>
<td>100(50%), 200(25%), 400(12.5%), 800(6.3%), 1600(3.1%), 3200(1.6%), 6400(0.8%), 12800(0.4%), 25600(0.2%), 51200(0.1%)</td>
</tr>
<tr>
<td>$C_{\text{band}}$</td>
<td>$T_{\text{comp}}$ x uniform[0.1, 0.5]</td>
</tr>
<tr>
<td>$C_{\text{lat}}$</td>
<td>$T_{\text{comp}}$ x uniform[0.1, 0.5]</td>
</tr>
<tr>
<td>Concurrency of Parallel Job</td>
<td>1(single), 2</td>
</tr>
<tr>
<td>Occurrence Rate of Parallel Job [%]</td>
<td>0, 1, 5, 10, 25, 50</td>
</tr>
<tr>
<td>Job Generation</td>
<td>120 jobs always exist</td>
</tr>
<tr>
<td>Number of Sites</td>
<td>3</td>
</tr>
<tr>
<td>Number of Nodes per Site</td>
<td>2</td>
</tr>
<tr>
<td>Scheduling Method</td>
<td>Round-Robin, History-based, Proposal</td>
</tr>
<tr>
<td>Total Simulation Time[sec.]</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>
## Simulation Parameters (2)

### Diagram

- **Site A**
- **Site B**
- **Site C**

### Table

<table>
<thead>
<tr>
<th>Connection</th>
<th>$F_{\text{band}}$</th>
<th>$F_{\text{lat}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-Site</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>A – B</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>B – C</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C – A</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
The Number of Executed Jobs

- Proposal improves the number of executed jobs

![Graph showing the number of executed jobs compared to different methods: Round-Robin, History-based, and Proposal. The Proposal method shows a higher number of executed jobs across different occurrence rates of parallel jobs.]
The idle time of nodes are caused by the startup synchronization of two processes assigned to different nodes.
The Distribution of the Waiting Time

**Average waiting time**

- Round-Robin
- History-based
- Proposal

**Maximum waiting time**

- Round-Robin
- History-based
- Proposal

The load-balancing by proposal realizes smallest variation in waiting time.
Future Plan

- The network configuration of Vector Computing Cloud is being updated.
  - Updating SINET3 to SINET4 (10Gbps)
  - Updating JGN2Plus to JGN-X (10Gbps, OpenFlow support)
  - There are multi path between Tohoku University and Osaka University.

- Plan of experiment
  - Evaluating the influence of bandwidth and latency on the communication time of various parallel applications.
  - Improving the accuracy of job execution time model
Conclusions

- Vector Computing Cloud
  - Vector Computing Cloud improves the flexibility of job execution and resource management
  - The wide-area vector co-operation has the potential for executing parallel job.

- Network-aware Job Scheduling
  - Modeling the execution time of parallel-job
  - Designing the resource allocation mechanism
  - The simulation results show that the Network-aware job scheduling realizes the appropriate resource allocation in Vector Computing cloud.
Thank you