Performance Studies of Scientific Applications on High Performance Computing Cluster with Big-Memory Footprint

By Derrick Lee, Garrett Lord, Jesse Martinez, and Christopher Moore

Mentors: HB Chen, Parks Fields, Benjamin McClelland, Alfred Torrez

Computer Systems, Cluster, and Networking Summer Institute 2011
Outline

- Big Memory Footprint
  - Introduction
  - Applications
  - Devices Used
  - Types of Benchmarks
  - Results from Benchmarks
  - Future Development
  - Conclusion
Introduction

- We have conducted a sequence of benchmark tests that are related to the LANL Data Intensive Super Computing (DISC) project.

- Virtualized Aggregation System

- The vSMP system will be well-suited to large memory problems.
Graph problems arise from a variety of sources

**Power Distribution Networks**

**Internet backbone**

**Social Networks**

**Graphs are everywhere!**

**Ground Transportation**

**Tree of Life**

**Protein-interaction networks**

Sources: C. Faloutsos talk [IPAM 05]

David A. Bader, Petascale Computing for Large-Scale Graph Problems
Certain activities are often suspicious not because of the characteristics of a single actor, but because of the interactions among a group of actors.

Interactions are modeled through a graph abstraction where there entities are represented by vertices, and their interactions are the directed edges in the graph.
Devices

- 4x IBM x3850 X5 servers
  - RAM: 256 GB per Node
  - Total: 4x 256 GB = 1TB
  - 128 Cores
  - ScaleMP Virtual Aggregation
  - vSMP Foundation (i.e. Versatile SMP)

- SGI UV 100/1000 system – 128 cores, 512GB memory, 2.67GHz/24MB E7- 8837 processors, remote testing account provided by SGI
Test Bed Setup

Infiniband QDR Switch

ScaleMP vSMP –
128 Cores
1024GB shared memory

IBM X3850 X5
32 Cores
256GB
memory
Two IB QDR
links

IBM X3850 X5
32 Cores
256GB
memory
Two IB QDR
links

IBM X3850 X5
32 Cores
256GB
memory
Two IB QDR
links

IBM X3850 X5
32 Cores
256GB
memory
Two IB QDR
links

Dual IB QDR links
AGGREGATION
Concatenation of physical resources

Virtual Machine

App

OS

ScaleMP™
How It Works

Multiple Computers with Multiple Operating Systems

Multiple Computers with a Single Operating System

InfiniBand Fabric

ScaleMP

APP

APP

APP

OS

InfiniBand Fabric

New Mexico CONSORTIUM

Los Alamos NATIONAL LABORATORY
Benchmarks

- STREAM Bandwidth
  - Copy, Scale, Add, Triad

- Graph500
  - Traversed Edges per Second (TEPS)
  - Scales 24-30
  - Cores: 64 and 128
STREAM Results - Triad

Rate (GB/s)

# of Cores

0 100 200 300

8 16 32 64 128
Graph 500 Results
Max. TEPS

<table>
<thead>
<tr>
<th>Scale Size</th>
<th>64 Cores</th>
<th>128 Cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale 24</td>
<td>1,500</td>
<td>2,000</td>
</tr>
<tr>
<td>Scale 25</td>
<td>1,200</td>
<td>1,800</td>
</tr>
<tr>
<td>Scale 26</td>
<td>2,500</td>
<td>3,000</td>
</tr>
<tr>
<td>Scale 27</td>
<td>1,800</td>
<td>2,300</td>
</tr>
<tr>
<td>Scale 28</td>
<td>2,000</td>
<td>2,500</td>
</tr>
<tr>
<td>Scale 29</td>
<td>1,300</td>
<td>1,900</td>
</tr>
<tr>
<td>Scale 30</td>
<td>1,000</td>
<td>1,500</td>
</tr>
</tbody>
</table>

*Test Incomplete
Graph 500 Results Compared to Current Placeholders

Rank

18  SuperDragon-1 (Sugon, 32 nodes / 384 cores) - ICT - Beijing, China
    1,454

19  Bigfoot (IBM x3850 x5, 4 nodes / 128 cores) - LANL
    1,036

20  Jaguar (Cray XT5-HE, 18,688 nodes / 224,256 cores) - ORNL
    1,011

21  Knot (HP MPI cluster, 8 processors / 64 cores) - USCB
    177

Max TEPS (Millions)
SGI UV 100 HW vs Big Memory Footprint SW - STREAM - MPI

- SGI – Triad
- Bigfoot – Triad

Threads vs Rate (GB/s)
Areas for Future Improvement

• Increasing RAM in each Node
• Adding SSDs for Swap File Systems
• Large Scale Graph Generation
• Large Data Analysis for Researchers
• Graph500 testing on SGI UV system
• Comparison of results from test beds
• Summit our Graph500 testing result from the scaleMP/IBM machine to Graph500 for SC11
Conclusion

• Large memory footprints provide performance increase for large data sets

• Graph 500 results showed potential increase for higher results.

• IBM aggregation system is more affordable compared to other machines for its current output
Questions?