Cluster OpenMP System: A Middleware for both Programmability and Performance

Jie Cai

Research Goal
The goal of this project is to provide a software platform that allows users using an easy programming model to achieve reasonable performance for some parallel applications on cluster systems.

Methodology
A two-step approach was used to achieve our research goal.
1) We have evaluated current Cluster-enabled OpenMP systems and find out the major system overhead is page/diff inter-node transfer caused by page fault servicing.
2) Two methods of reducing the cost of page/diff inter-node transfer.
   (a) Utilizing multiple network connections to reduce single data transfer time.
   (b) Developing and implementing prefetch techniques to amortize network latencies and leverage multiple network connections to achieve higher throughput.

Research Outcomes
We analyzed and modeled performance of Cluster OpenMP systems. Multirail network technique was investigated based on InfiniBand interconnects. Three prefetch techniques have been developed and implemented. Data are successfully aggregated, network latencies are amortized, and it shows great potential to leverage high throughput of multirail networks. Our research goal is partially completed. We believe that, with our contribution, even KIDs can do cluster programming!

Supervised by Dr. Peter Strazdins and Dr. Alistair Rendell, in co-operation with Intel.

Scenario I – Nothing is Perfect.
- Poor Performance on small message size!

Scenario II - A simple solution is limited.
- Design of multiple physical network connections.
- Multiple connections only obtain better bandwidth for message size larger than 4KB.

Scenario III - A hybrid solution better solves the problem.
- We developed three different prefetch techniques, and are approaching to the end of implementation phase. Some preliminary results shown below to demonstrate how to achieve the “double wins”.

The first win: reduced number of message communicated.

The second win: larger average message size!

Remark: the average message size may be vary between different programs.