Using *BeoSim* to Evaluate Bandwidth-aware Meta-schedulers for Co-allocating Jobs in a Mini-grid*

Parallel Architecture Research Lab
Clemson University
http://www.parl.clemson.edu

Clusters of commodity processors have become fixtures in research laboratories around the world. Collections of several co-located clusters exist in many larger laboratories, universities, and research parks. This co-location of several resource collections naturally lends itself to the formation of a mini-grid.

A mini-grid is distinguished from a traditional computational grid in that the mini-grid utilizes a dedicated interconnection network between grid resources with a known topology and predictable performance characteristics. This type of networking infrastructure allows for the possibility of mapping jobs across cluster boundaries in a process known as *co-allocation* or *multi-site scheduling* (Figure 1).

In this research, we develop a parallel job model that takes both computation and communication into account as a means by which to explore co-allocating grid schedulers that exploit these unique architectural features.

We also develop several bandwidth-aware co-allocating meta-schedulers that take inter-cluster network utilization into account as a means by which to mitigate the slowdown associated with the interaction of simultaneously co-allocated jobs in a mini-grid. By making use of a bandwidth-centric parallel job communication model that captures the time-varying utilization of shared inter-cluster network resources, we are able to evaluate the performance of grid scheduling algorithms that focus not only on computational resource allocation, but also on shared inter-cluster network bandwidth.

Previous work in the area of job co-allocation tends to characterize jobs by either specifying that all communications require a fixed amount of time to travel between clusters or by assigning co-allocated jobs a fixed execution-time penalty. This type of characterization is not sensitive to the time-varying contention for bandwidth in the inter-cluster communication links and the impact it has on the execution time of co-allocated jobs that share network resources. We take a different approach by considering that as jobs become co-allocated or co-allocated jobs terminate, there is a continual change in the available inter-cluster bandwidth. Therefore, in our work, the duration of wide area communication is a function of the time-varying network bandwidth utilization among clusters participating in the mini-grid, which in turn affects the execution time of co-allocated jobs. This research aims to extend the previous efforts by replacing the static communication model with a more dynamic view of job communication that is *bandwidth-centric*.

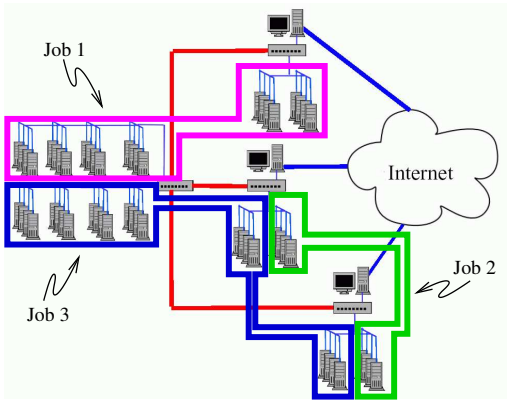

Figure 1. Job co-allocation

We find that schedulers designed to allocate node resources across cluster boundaries can result in rather poor overall performance over a wide range of workload characterizations and mini-grid configurations due to the interaction simultaneously co-allocated jobs experience as they contend for inter-cluster network bandwidth. Our research therefore focuses on a range of algorithms with varying levels of complexity that attempt to mitigate this impact.

We make use a discrete event-driven simulator known as BeoSim to evaluate the performance of the parallel job scheduling algorithms. To learn more about this research, please feel free to visit our website at **http://www.parl.clemson.edu/beosim**.

*This is a handout for the SC2004 NASA booth.