SCHEDULING PARALLEL JOBS IN A MINI-GRID

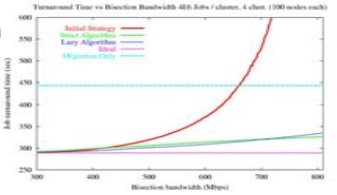
William M. Jones,
Louis W. Pang, Walter B. Ligon, III
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
CLEMSON UNIVERSITY

Job Scheduling

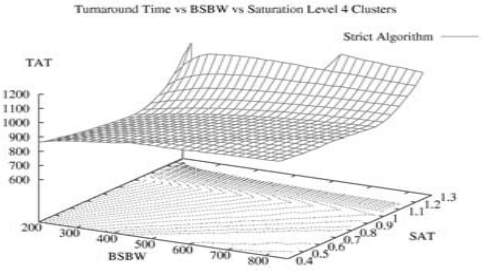
- Resource allocation in space-time
  - Where to run job?
  - When to run job?
- Job migration
  - Moving entire job to a remote cluster
- Job co-allocation
  - Mapping jobs across cluster boundaries
  - Node resource allocation
  - Network bandwidth contention
- Focus: Investigate these issues with a grid simulator

Initial Findings

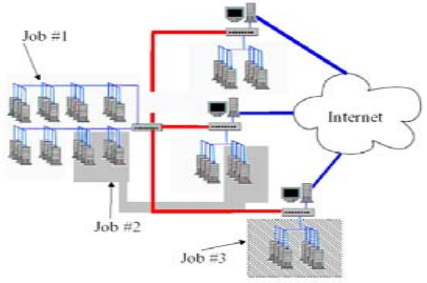
- Ideal case
  - Unlimited bandwidth
  - No penalty
  - Lower bound
- Migration only
  - Upper bound
- New Algorithms
- Bandwidth-aware
- Practical



Strict Algorithm Performance



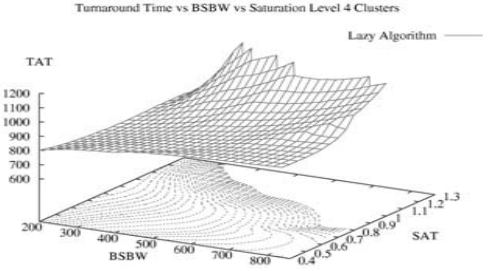
Job Co-allocation



Bandwidth-aware Schedulers

- Inter-cluster bandwidth saturation
- Communication slowdown
- Sensitive to network contention
- Strategy
  - Attempt to allocate resources locally
  - Attempt to migrate entire job
  - Attempt to co-allocate job subject to network load
- Mitigating impact to inter-cluster network saturation

Lazy Algorithm Performance

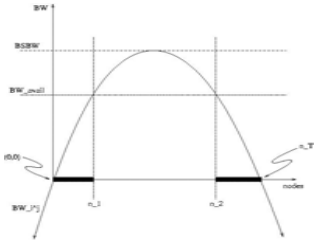


Models

- Communication
  - All-to-all collective communication patterns
  - Bisection bandwidth requirements
- Parallel job execution
  - Computation + communication time
  - Communication time subject to network congestion
- Mini-grid architecture
  - Homogenous clusters
  - Dedicated interconnection network

Strict Algorithm

- Communication model
- Possible job partitions
- Constraint satisfaction
- Optimization
- Requires foreknowledge
- Computationally expensive
- Practical?
- Better algorithms?



Future Work

- Integrate local cluster policy into scheduling
- Improve job communication model
- Verify saturation model with empirical data
- Extend network to include additional topologies
- Generate more realistic workloads
- Design more intelligent scheduling algorithms
- Create more extensive priority mechanisms
- Implement actual scheduler for our mini-grid

Acknowledgements: *This work was supported by the ERC program of the National Science Foundation under Award Number EEC-9731680.*

CAEFF