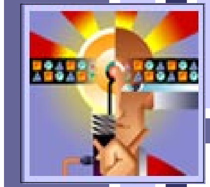


Graph Optimization Algorithms for Sun Grid Engine

Lev Markov

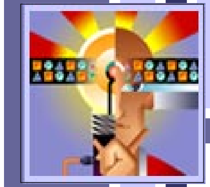
Sun Grid Engine

- SGE – management software that optimizes utilization of software and hardware resources in heterogeneous networked environment.
- SGE – distributes computational workload simultaneously increasing productivity of machines.
- SGE – maximizes the number of completed jobs.

Job Scheduling and Assignment within Sun Grid Engine

- Goal:
  - Select processing resource for every job.
  - Select job processing order for every resource.
- Constraints:
  - data/time dependencies between jobs.
  - limitation of data links between resources.
  - processing limitation of resources.
  - individual requirements of jobs.
  - ...

Job Scheduling and Assignment within Sun Grid Engine



- New Features:
  - Data/time dependencies between jobs.
  - Data communication links between resources.
  - Job deadlines.
  - Job preemption.
  - Advance reservation.
  - Automated global job priorities to guide the entire scheduling and assignment process.

Technical Challenges

- Deal with REAL networked resource management problem.
  - all required constraints
  - all required scheduling features
- Global approach vs. manual priorities.
- Speed of the algorithms.

Input Data

- Properties of jobs
- Properties of resources
- Relations between jobs and processing resources
- Optimization parameters
- Required scheduling features

Properties of Jobs

- Initial priority
- Dependence on other jobs (time or data)
- Allowed types of resources
- Required licenses
- Permission to partition into parallel sub-jobs
- Permission to preempt
- Permission to restart
- Completion deadline

Properties of Resources

- Resource hierarchy
- Hierarchical allocation of processing slots
- Hierarchical memory allocation
- Hierarchical allocation of licenses
- Hierarchical allocation of user defined resources
- Link bandwidth between resources

Relations between Jobs and Processing Resources

- Processing speed
- Required memory
- Required number of processing slots

Optimization Parameters

- **Parameters controlling job priorities**
  - Importance of required memory
  - Importance of required processing slots
  - Importance of available time slack
  - Importance of initial priorities
  - Importance of waiting time

Optimization Parameters

- **Parameters controlling preemption strategy**
  - (Time required to finish) / (Time already received) – **controls preemption of a job**
  - (Time before preemption) / (Total execution time) – **controls start of a job**
  - Ratio between two job priorities – **controls a possibility of preemption by a job**

Required Scheduling Features

- Automatic partitioning of large parallel jobs
- Automatic scheduling around pre-assigned jobs
- Automatic advance reservation
- Automatic job back filling
- Automatic job preemption

Technical Approach

- Directed graph (job graph)
  - Job properties attached to the nodes
  - Link weights deal with time delay and/or data flow
- Non-directed graph (resource graph)
  - Resource properties attached to the nodes
  - Link weights deal with quality of communication channels
- Job graph nodes are associated with parts of the resource graph

Technical Approach

- Two stage optimization process
  - First stage (one path):
    - Job graph nodes get global static priorities
    - Jobs are selected based on static priorities
  - Second stage (one path for every job node):
    - Resource graph nodes get global dynamic priorities
    - Resources are selected based on dynamic priorities

Scheduling Features

- Data/time dependent jobs.
- Preemption of low priority jobs.
- Advance reservation of high priority jobs.
- Job deadlines.
- Automatic partitioning of large parallel jobs.

Performance Results

UltraSPARC II @450Mhz

(non-optimized code compiled with debug option)

# of Jobs	# of Resources	CPU time
150	4	0.09 sec
230	4	0.15 sec
6700	7	40.00 sec

Technical Status

- The first version of a prototype system is finished and transferred to the SGE group.
- All advance scheduling and assignment features are in place.
- Some new scheduling features will be used starting with 6.0 release of SGE EE.
- Speed requirements for the new algorithms are satisfied.