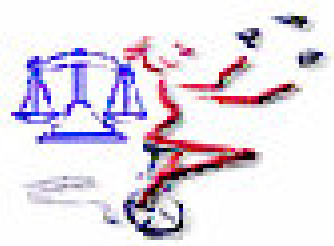
the gridbus project

Libra: An Economy driven Job Scheduling System for Clusters



Jahanzeb Sherwani[1], Nosheen Ali[1], Nausheen Lotia[1], Zahra Hayat[1], **Rajkumar Buyya**[2]

1. Lahore University of Science and Management (LUMS), Lahore, Pakistan

2. Grid Computing and Distributed Systems (GRIDS) Lab., University of Melbourne, Australia

www.gridbus.org

THE UNIVERSITY OF MELBOURNE

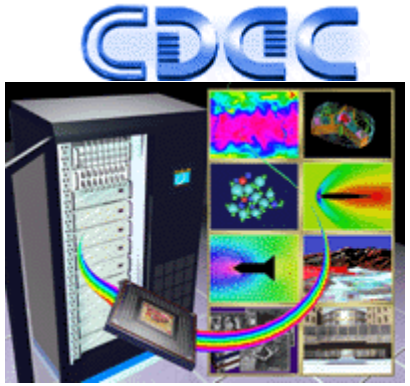
Agenda

- Introduction/Motivations
- The Libra Scheduler Architecture & Cost-based Scheduling Strategy
- Implementation
- Performance Evaluation
- Conclusion and Future Work

Introduction

- Clusters (of "commodity" computers) have emerged as mainstream parallel and distributed platforms for high performance, high-throughput and high-availability computing.
- They have been used in solving numerous problems in science, engineering, and commerce.

Adoption of the Approach



HEWLETT PACKARD
Expanding Possibilities
HP NetServer Cluster Solutions

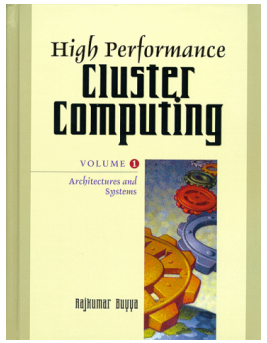
HOTBOT

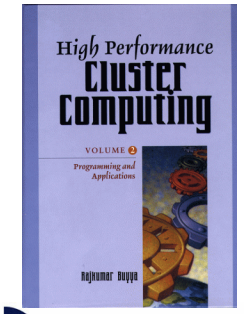
Sun microsystems

Oracle

Buy it now!
amazon.com

hotmail™

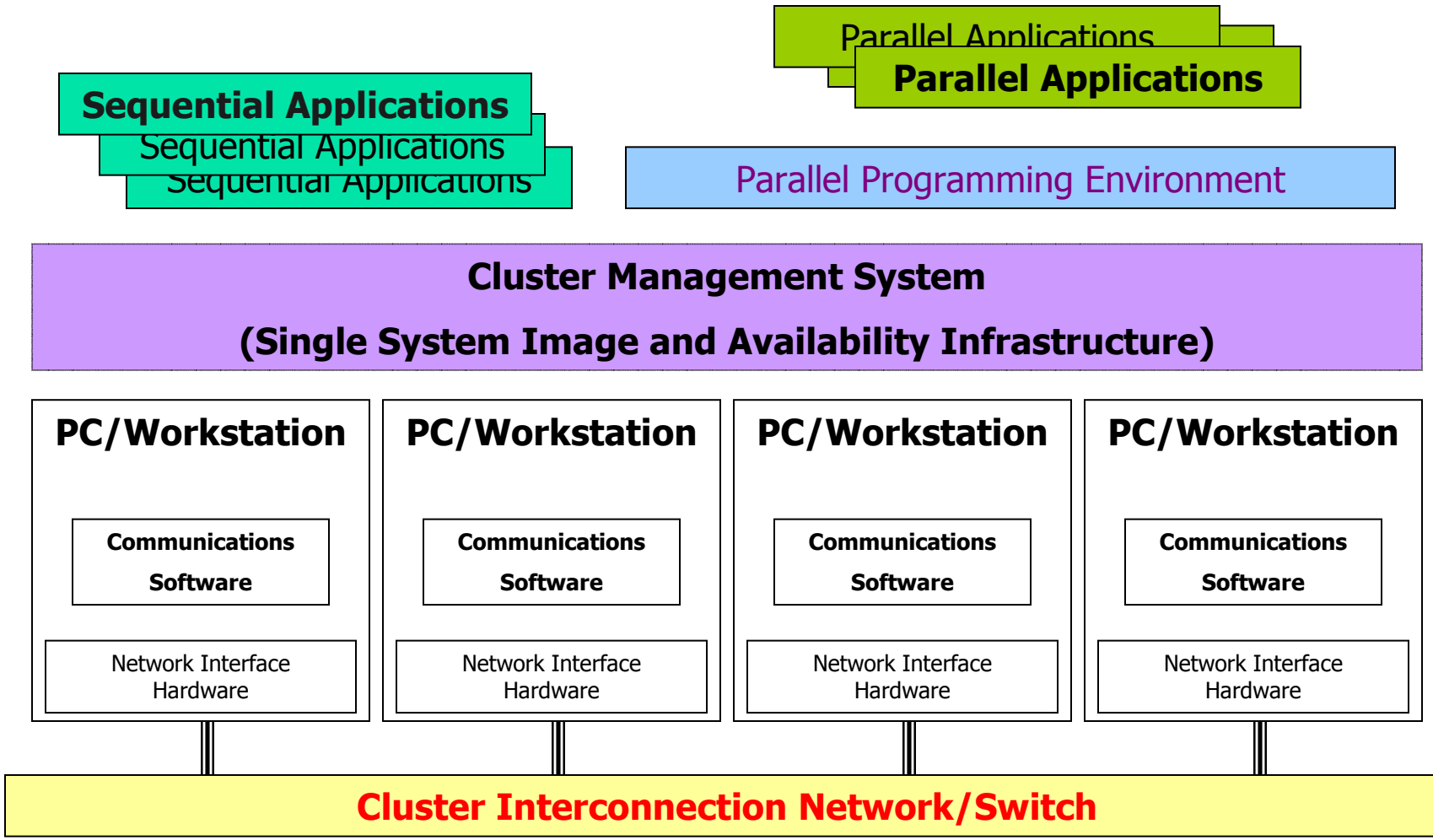




Google™

Microsoft®

Cluster Resource Management System: Managing the Shared Facility



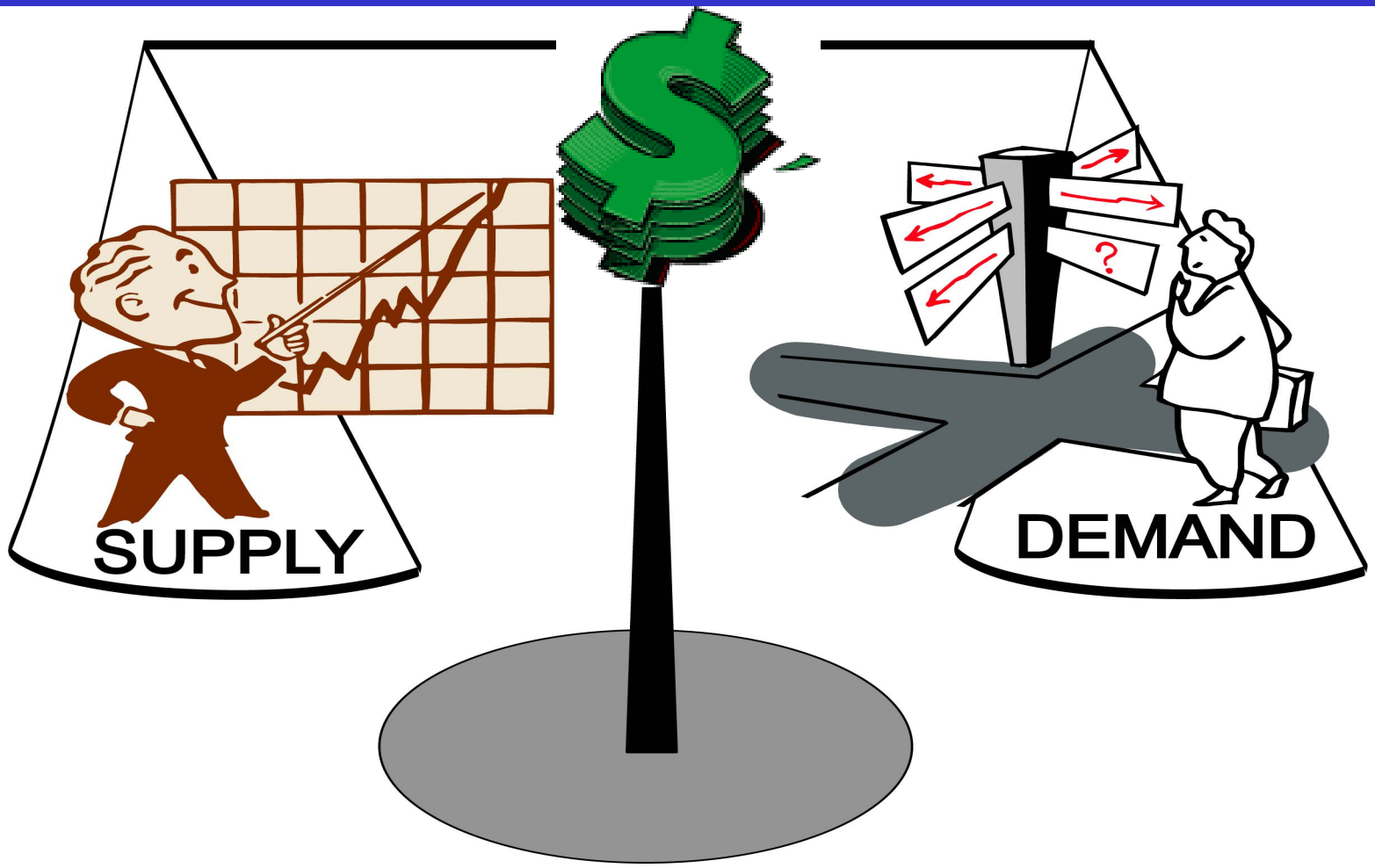
Some Cluster Management Systems

- Commercial and Open-source Cluster Management Software
- Open-source Cluster Management Software
  - DQS (Distributed Queuing System)
  - Condor
  - GNQS (Generalized Network Queuing System)
  - MOSIX
  - Load Leveler
  - **SGE (Sun Grid Engine)**
  - **PBS (Portable Batch System)**

Cluster Management Systems Still Use System Centric Approach

- Traditional CMSs focus has essentially been on maximizing CPU performance, but not on improving the value of utility delivered to the user and quality of services.
- Traditional system-centric performance metrics
  - CPU Throughput
  - Mean Response Time
  - Shortest Job First
  - FCFS
  - Some Static Priorities
  - ...

The Libra Approach: Computational Economy Paradigm for Management & Job Scheduling

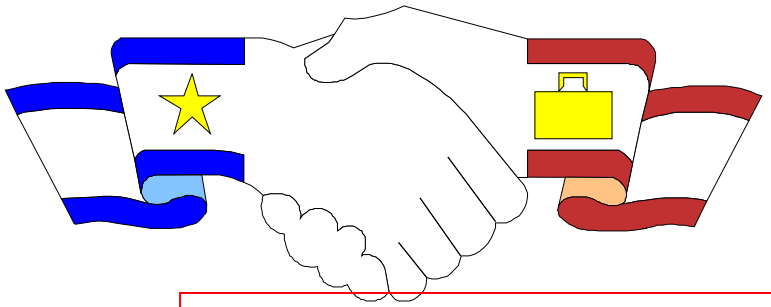


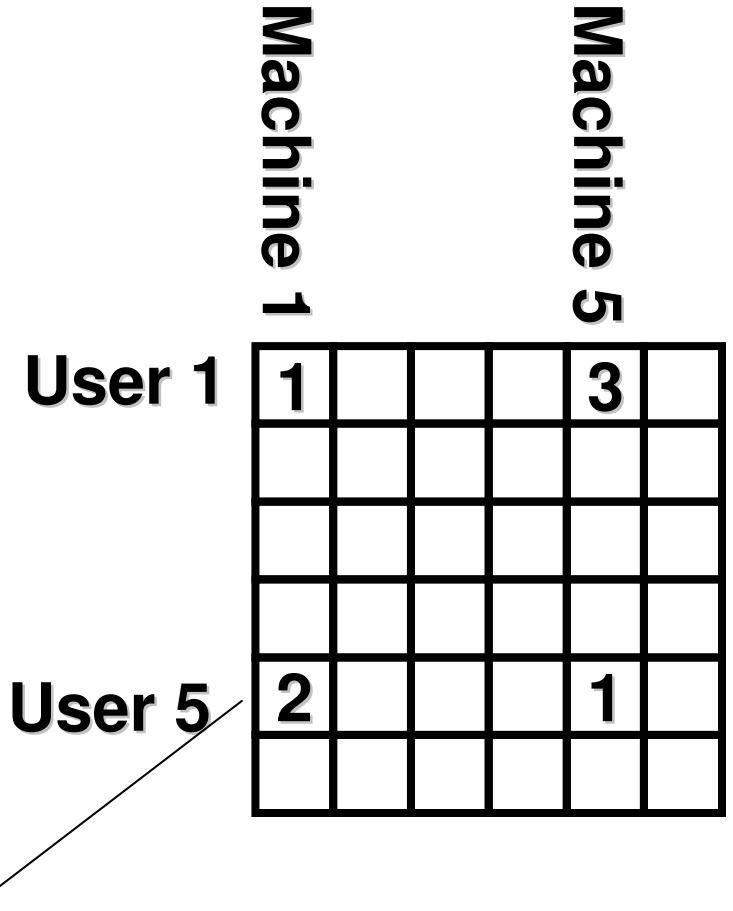
Cost Model: Why are they needed ?

- Without cost model any shared system becomes un-manageable
- It supports QoS based resource allocation and help manage supply-and-demand for resources.
- Improves the value of utility delivered.
- Also, improves the resource utilization.
- Cost units (G$) may be
  - Rupees/Dollars (real money)
  - Shares in global facility
  - Stored in bank

Cost Matrix

- Non-uniform costing
  - Different users are charged different prices that vary with time.





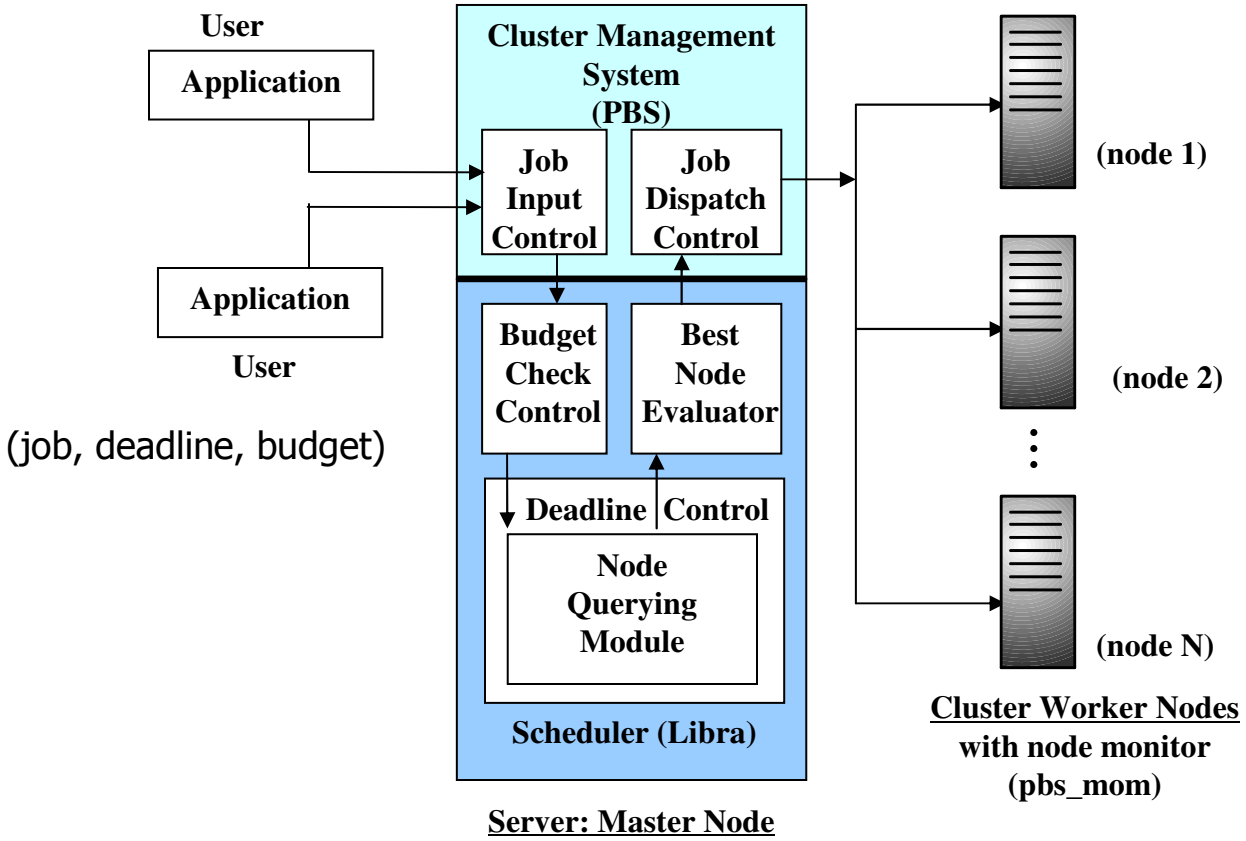
Resource Cost = Function (cpu, memory, disk, network, software, QoS, current demand, etc.)

Simple: price based on peaktime, offpeak, discount when less demand, ..

Computational Economy Parameters

- Job parameters most relevant to user-centric scheduling
  - Budget allocated to job by user
  - Deadline specified by user

Libra Architecture



Libra with PBS

- Portable Batch System (PBS) as the Cluster Management Software (CMS)
  - Robust, portable, effective, extensible batch job queuing and resource management system
  - Supports different schedulers
  - Job accounting
  - Allows Plugging of Third-Party Scheduling Solution

The Libra Scheduler

- Job Input Controller
  - Adding parameters at job submission time
    - *deadline*
    - *budget*
    - *Execution Time*
  - Defining new attributes of job
- Job Acceptance and Assignment Controller
  - Budget checked through cost function
  - Admission control through deadline scheduling
  - Execution host with the minimum load and ability to finish job on time selected
  - Node Resource Share Allocation: Proportional to the needs of multiple User Jobs QoS needs.

The Libra Scheduler

- Job Execution Controller
  - Job run on the best node according to algorithm
  - Cluster and node status updated
    - *runTime*
    - *cpuLoad*
- Job Querying Controller
  - Server, Scheduler, Exec Host, and Accounting Logs

Pricing the Cluster Resources

- Cost= α * (Job Execution Time) + β * (Job Execution Time / Deadline)
  - Cost = α *E + β *E/D (where α and β are coefficients)
- Cost of using the cluster depends on job length and job deadline: the longer the user is prepared to wait for the results, the lower his cost
- Cost formula motivates users to reveal their true QoS requirements (e.g., deadline)

PBS-Libra Web --- Front-end for the Libra Engine

PBS-Libra Web Login Page

Login

User Name:

Password:

Submit

Welcome to the Web front-end of Libra -- an Economy-Driven Cluster Scheduler. The Libra team is:

Project Owner/Client: Rajkumar Buyya (rajkumar@csse.monash.edu.au)

Jahanzeb Sherwani (2002-02-0058)
Nosheen Ali (2002-02-0113)
Nausheen Lotia (2002-02-0111)
Zahra Hayat (2002-02-0189)

Lahore University of Management Sciences

Department of Computer Science

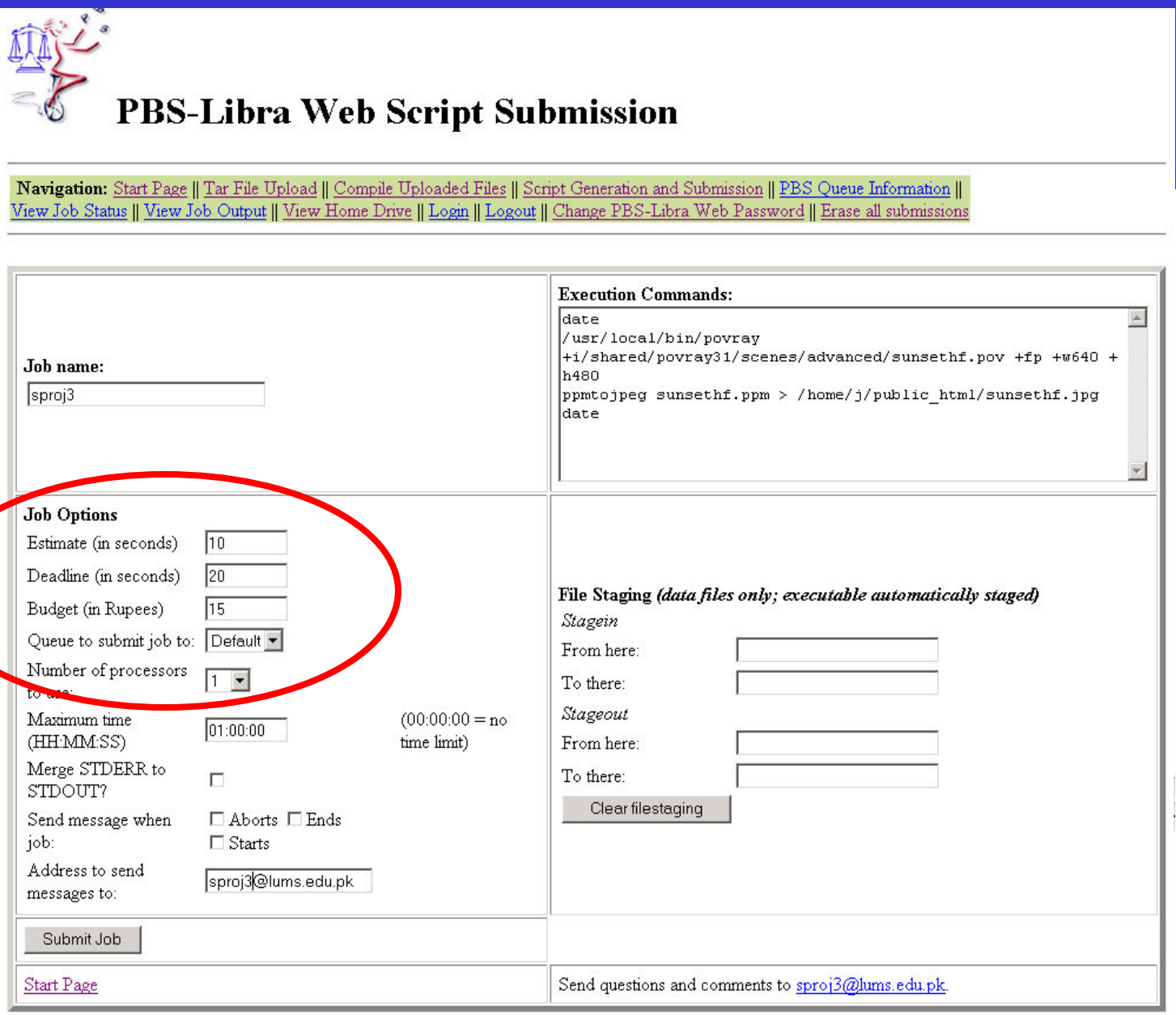
in collaboration with

MONASH UNIVERSITY

School of Computer Science and Software Engineering

Send questions and comments to sproj3@lums.edu.pk. You can find help here .

PBS-Libra Web



PBS-Libra Web

Job Status for Job 393

Navigation: Start Page || Tar File Upload || Compile Uploaded Files || Script Generation and Submission || PBS Queue Information || View Job Status || View Job Output || View Home Drive || Login || Logout || Change PBS-Libra Web Password || Erase all submissions

```
Job Id: 393.mspc37.lums.edu.pk
    Job_Name = sproj3
    Job_Owner = j@mspc37.lums.edu.pk
    resources_used.cput = 00:00:00
    resources_used.mem = 2856kb
    resources_used.vmem = 6484kb
    resources_used.walltime = 00:00:00
    job_state = R
    queue = dque
    server = mspc37.lums.edu.pk
    Checkpoint = u
    ctime = Thu May  9 02:29:51 2002
    Error_Path = mspc37.lums.edu.pk:/home/j/pbsweb/libra/sproj3.e393
    exec_host = mspc37/0
    Hold_Types = n
    Join_Path = n
    Keep_Files = n
    Mail_Points = aeb
    Mail_Users = sproj3@lums.edu.pk
    mtime = Thu May  9 02:29:51 2002
    Output_Path = mspc37.lums.edu.pk:/home/j/pbsweb/libra/sproj3.o393
    Priority = 0
    qtime = Thu May  9 02:29:51 2002
    Rerunable = True
    Resource_List.ncpus = 1
    Resource_List.walltime = 01:00:00
    session_id = 28394
    Shell_Path_List = /bin/sh
    Variable_List = PBS_O_HOME=/home/j,PBS_O_LOGNAME=j,
        PBS_O_PATH=/usr/local/bin:/bin:/usr/bin:/shared/pvm3/lib,
        PBS_O_MAIL=/var/mail/j,PBS_O_SHELL=/bin/bash,
        PBS_O_HOST=mspc37.lums.edu.pk,PBS_O_WORKDIR=/home/j/pbsweb/libra,
        PBS_O_QUEUE=dque
    etime = Thu May  9 02:29:51 2002
    budget = 15
    deadline = 20
    estimate = 10
```

Performance Evaluation: Simulations

- Goal:
  - Measure the performance of Libra Scheduler
- Performance = ?
  - Maximize user satisfaction
  - Maximise value delivered by the utility
- Simulation Platform: GridSim
  - Simulated scheduling using the GridSim toolkit
  - http://www.gridbus.org/gridsim

Simulations

- Methodology
  - Workload
    - 120 jobs with deadlines and budgets
    - Job lengths: 1000 to 10000 (MIs)
  - Resources
    - 10 node, single processor (MIPS rating: 100) (homogenous) cluster

Simulations

- Scheduler simulated as a function
  - Input: job size, deadline, budget
  - Output: accept/reject, node #, share allocated

Simulations

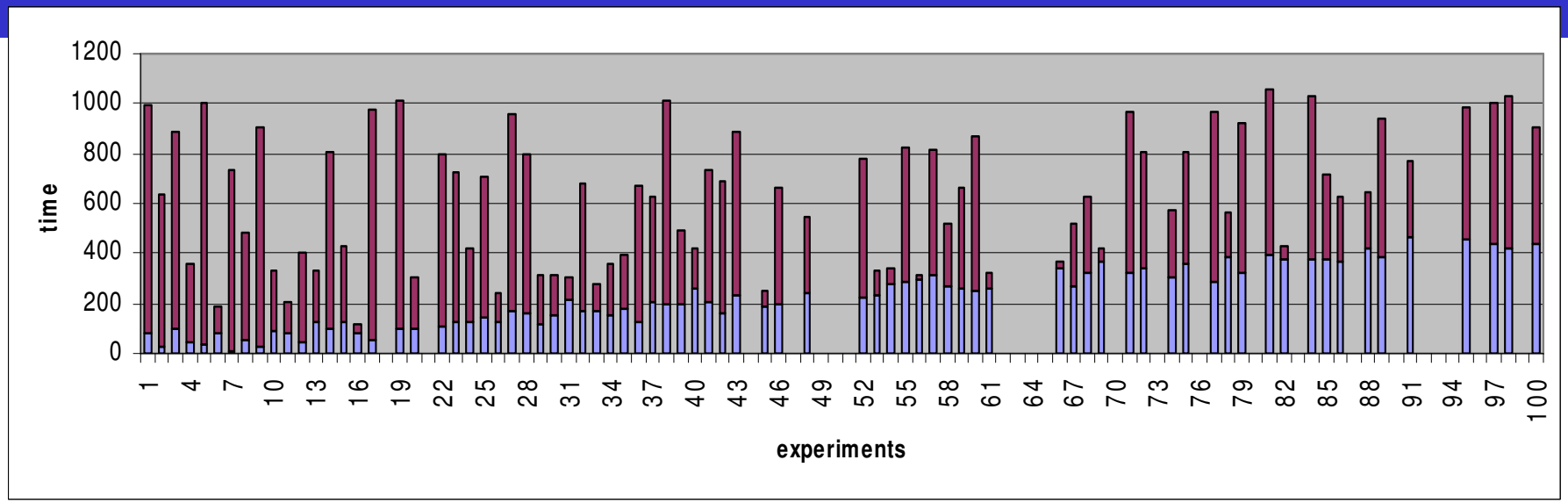
- Compared:
  - Proportional Share (Libra)
  - FIFO (PBS)
- Experiments:
  - 120 jobs, 10 nodes
  - Increasing workload to 150 and 200
  - Increasing cluster size to 20

Simulation Results

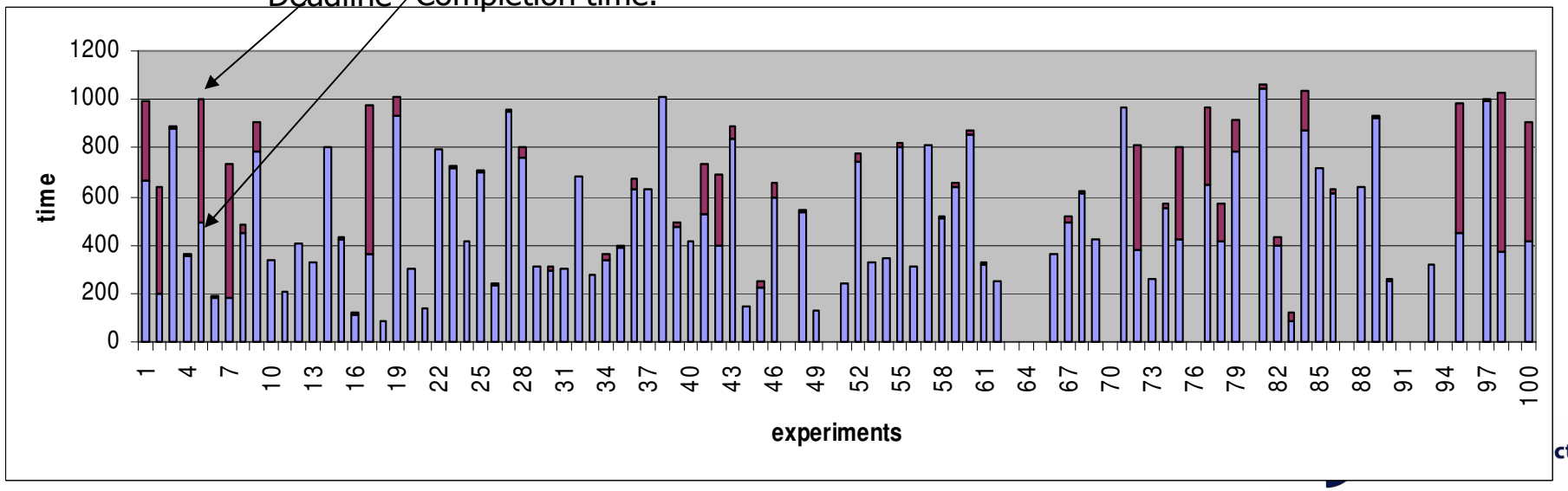
- 120 jobs, 20 did not meet budget

100 Jobs, 10 Nodes
FIFO: 23 rejected - Proportional Share: 14 rejected

PBS FIFO



Libra Proportional

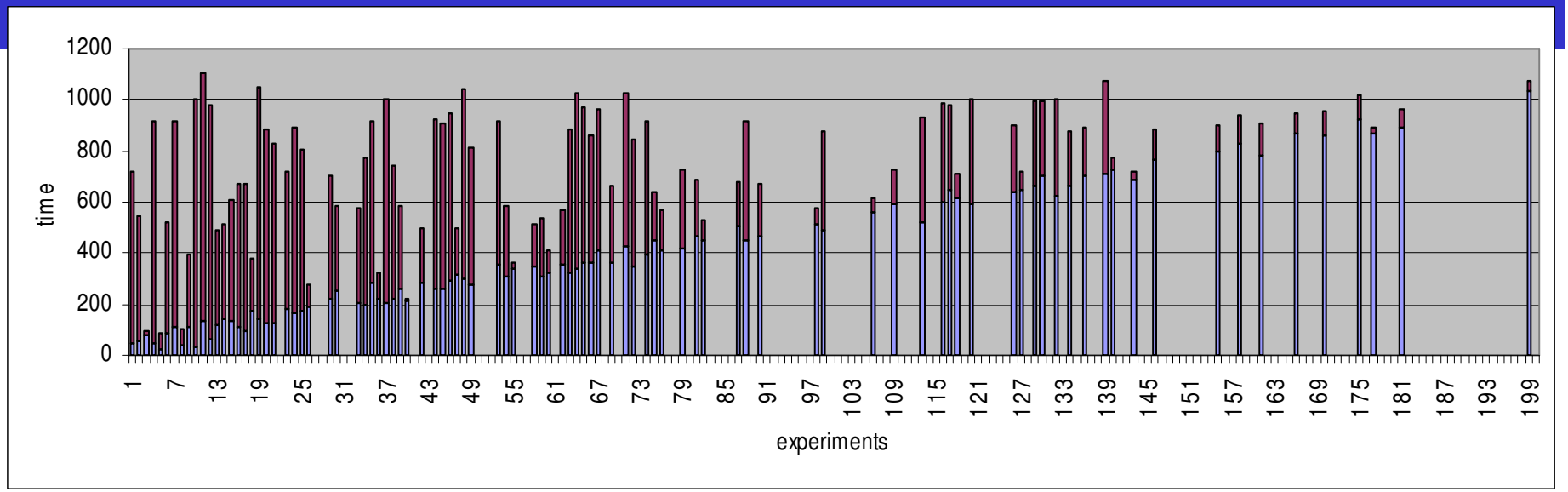


Simulation Results

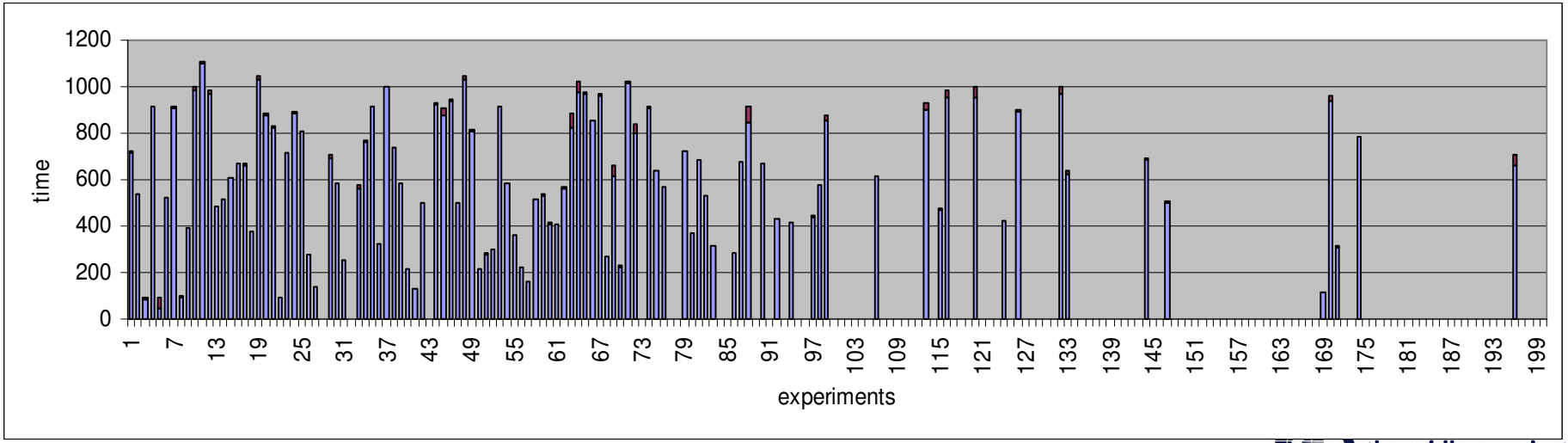
- Increase workload to 200 jobs on the same 10 node cluster

200 Jobs, 10 Nodes
FIFO: 105 rejected - Proportional Share: 93 rejected

PBS FIFO



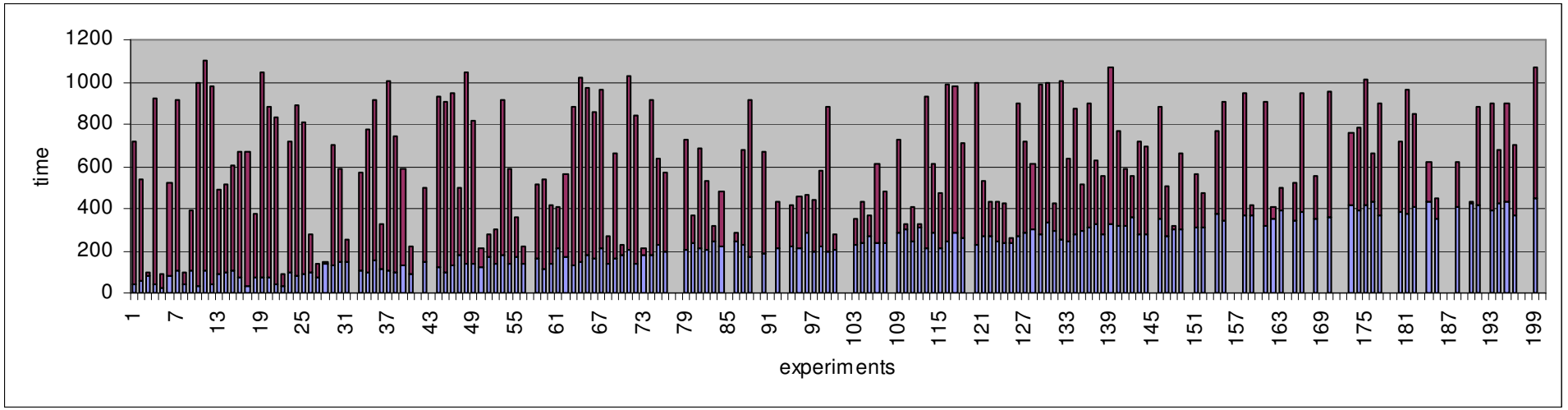
Libra Proportional

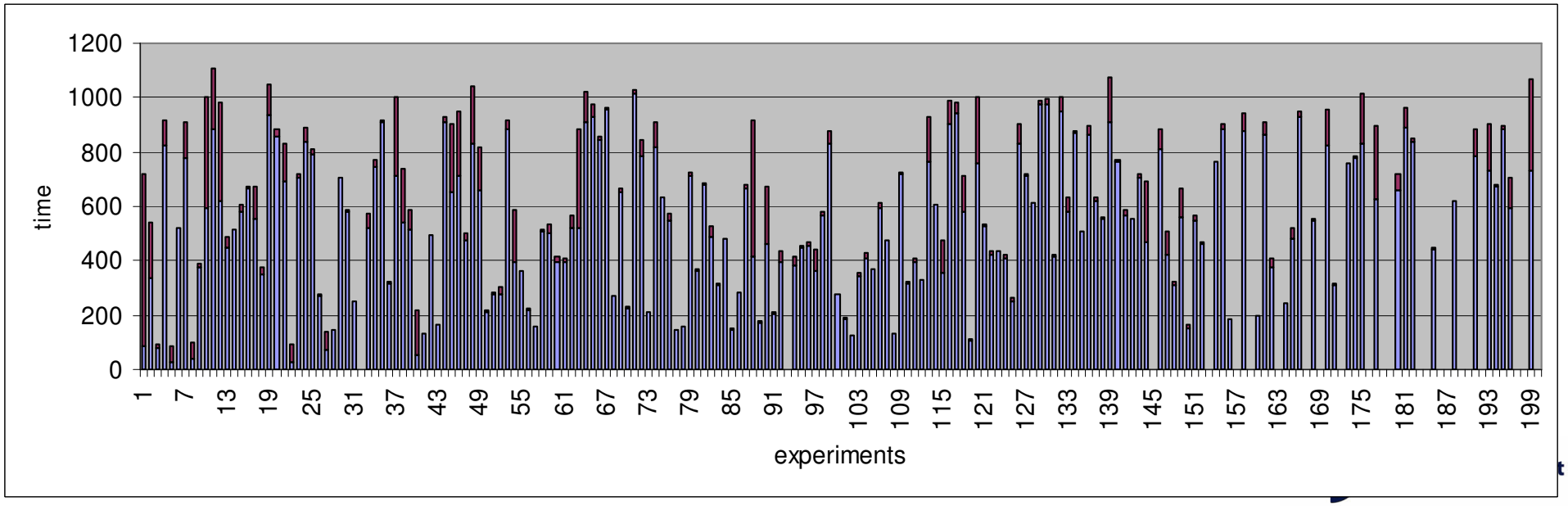


Simulation Results

- Scale the cluster up to 20 nodes

200 Jobs, 20 Nodes
FIFO: 35 rejected - Proportional Share: 23 rejected





PBS FIFO & Libra Strategy

No. of Jobs	No. of Nodes	No. of Jobs Accepted		No. of Jobs Rejected	
		PBS FIFO	Libra	PBS FIFO	Libra
100	10	77	86	23	14
	20	90	86	10	14
200	10	95	102	105	98
	20	165	177	35	23

Conclusion & Future Work

- Successfully developed a Linux-based cluster that schedules jobs using PBS with our economy-driven Libra scheduler, and PBS-Libra Web as the front end.
- Successfully tested our scheduling policy
- Proportional Share delivers more value to users
- Exploring other pricing mechanisms
- Expanding the cluster with more nodes and with support for parallel jobs
- Implement Libra for SGE (Sun Grid Engine)
  - Sponsored by Sun!

Thank you