IRIX Resource Management Plans & Status

Dan Higgins

djh@sgi.com

Engineering Manager
Resource Management Team

SGI

41st Cray User Group Conference
Minneapolis, Minnesota

IRIX Resource Management

Overview

- IRIX Job Limits
- IRIX Comprehensive System Accounting (CSA)
- IRIX Scheduling
  - Share II Fair Share Scheduler
  - Miser
  - eXtensible Resource Scheduler (XRS)
- Workload management
  - LSF Integration
  - NQE

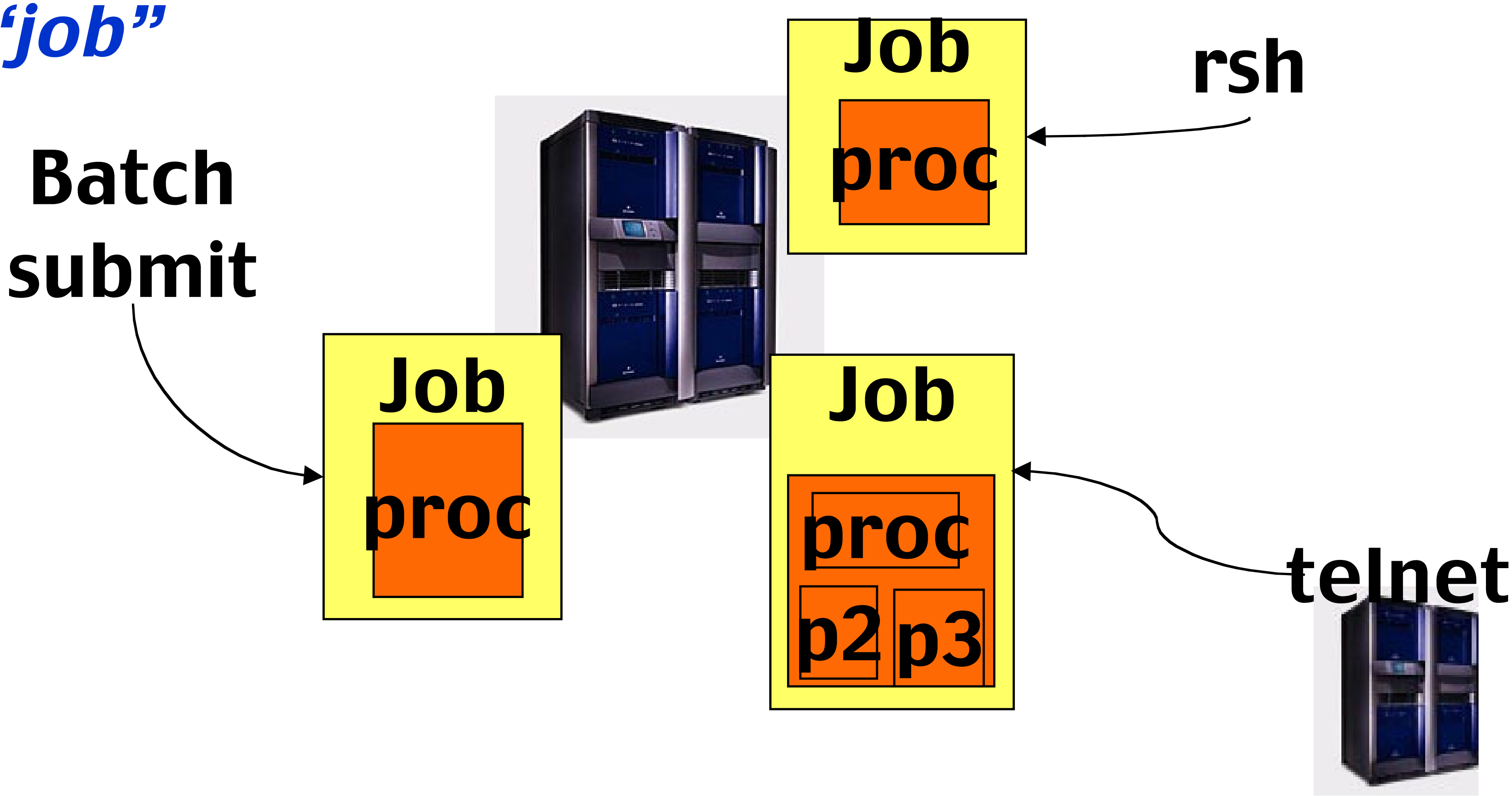
IRIX Job Limits

What is it?

- Job Concept
- Limit Domains
- Supported Limits

IRIX Job Concept

Every connection to the machine starts a "job"



Batch submit

Job

proc

rsh

Job

proc

telnet

Job

proc

p2

p3

Limit Domains

- Allows administrators & vendors to set limits on a per-user basis
- Extendable domains - batch, interactive, ++
- Limits set when a job is initiated

Supported Limits for Jobs

- Extends current IRIX process limits across all processes within a job
- A couple new job-only limits to limit number of processes and tapes (enforceable by TMF) per job
- Used via new setjlimit(2) & getjlimit(2) calls
- jlimit command displays or alters job limits
- Ps command modified to show job ids
- Job ids are unique in a cluster

IRIX Job Limits

Status

- Requirements, User Interface and Design documents are complete
- Much of the IRIX kernel changes are complete
- Beta testing in September at Boeing
- Generally availability with IRIX 6.5.7 in Q1CY00
- Integrating IRIX Job Limits with LSF

IRIX Comprehensive System Accounting (CSA)

An alternative accounting package for customers that demand more detail

- Use Cray accounting functionality with IRIX terminology
- Standard UNIX V accounting and IRIX extended accounting still supported and coexist
- Published API for vendor integration

IRIX CSA Features

Phase 1

- Per-job accounting
- User job accounting (ja command)
- Daemon accounting
- Flexible accounting periods
- Flexible system billing units (SBUs)
- +++

IRIX Comprehensive System Accounting (CSA)

Status

- Requirements and Design documents complete
- Significant amount of coding for IRIX kernel changes already complete
- Beta testing in December at Boeing
- General availability with IRIX 6.5.8 Q2CY00
- Integrating IRIX CSA with LSF

IRIX CSA Futures

sgi

Features for consideration (post phase 1)

- Support for specific hardware capabilities:
  - Multi-tasking records
  - MPP records for MPI jobs
- Incremental accounting for long running jobs
- Accounting by Array Session Handle (ASH)
- API for reading the accounting records

IRIX Scheduling

Overview

- Share II
- Miser
- eXtensible resource scheduler (XRS)

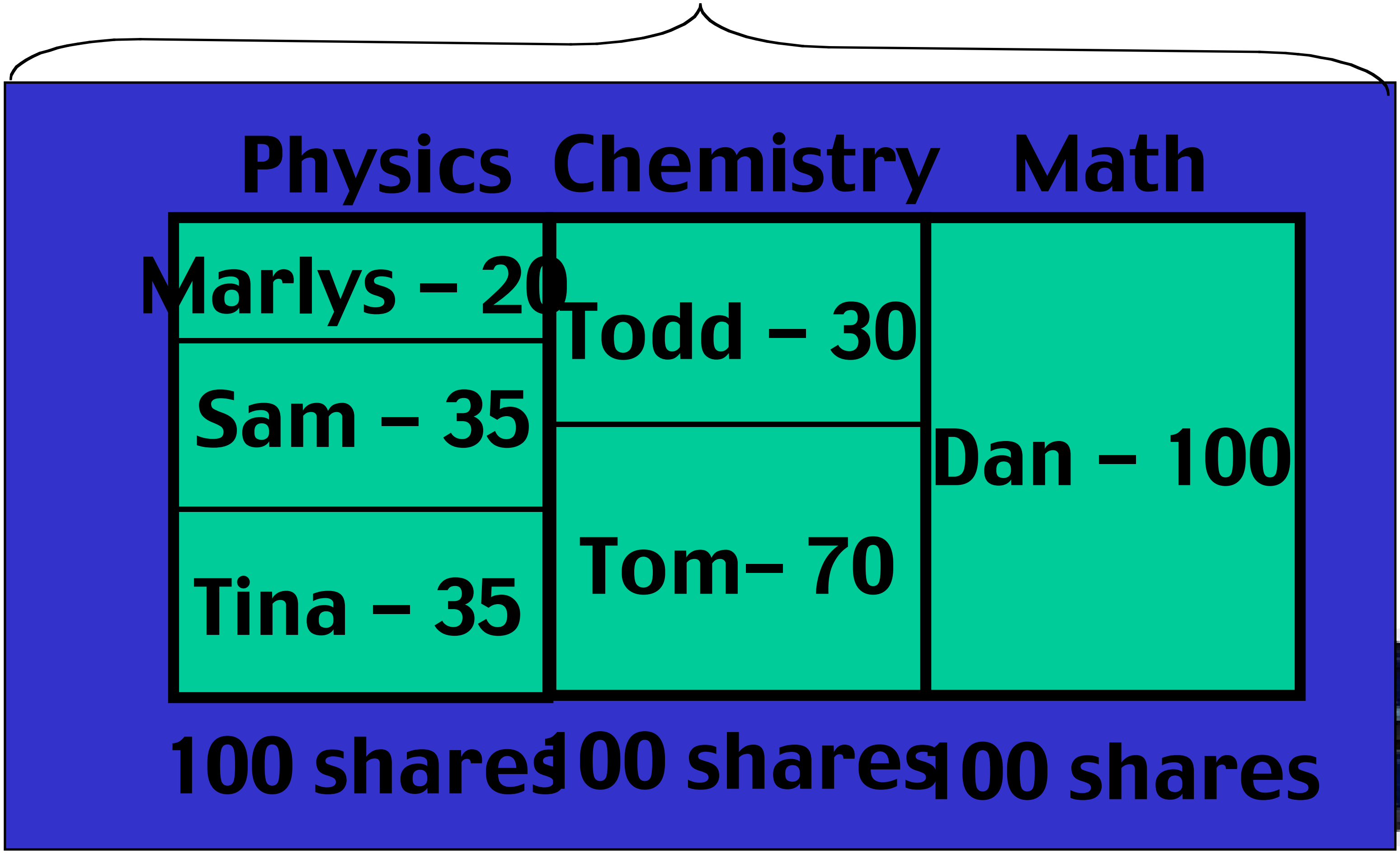
Share II Resource Manager

“Fair share” scheduling

- Users and/or Groups can be guaranteed a certain percentage of the machine
- Uses group dynamics to keep overall usage fair
- Often used when multiple groups share machine
- Currently single system only
- Available for IRIX 6.5

Share II Resource Manager

Single system Origin



Physics

Marlys – 20

Sam – 35

Tina – 35

100 shares

Chemistry

Todd – 30

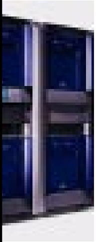
Tom– 70

100 shares

Math

Dan – 100

100 shares



Miser

Overview

- Deterministic batch scheduler for applications with known time and space requirements
- Generally Available since IRIX 6.5
- Didn’t quite meet user’s functional expectations
- Had some stability issues

Miser

Many improvements

- Improved Repeatability
- Many Miser related panics fixed
- Added repack policy (backfill)
- Increased performance & CPU utilization
- Miser_cpuset job tracking problem
- Miser_cpuset recovery mechanism
- Additional information in command output
- Better documentation

Miser

Plans

- Evaluating Integration of miser Q's & miser_cpusets
- Integrating Miser & miser_cpusets with LSF 4.0 (Available Q4CY99)
- Fix critical customer issues
- Add new functionality into XRS

IRIX eXtensible Resource Scheduler (XRS)

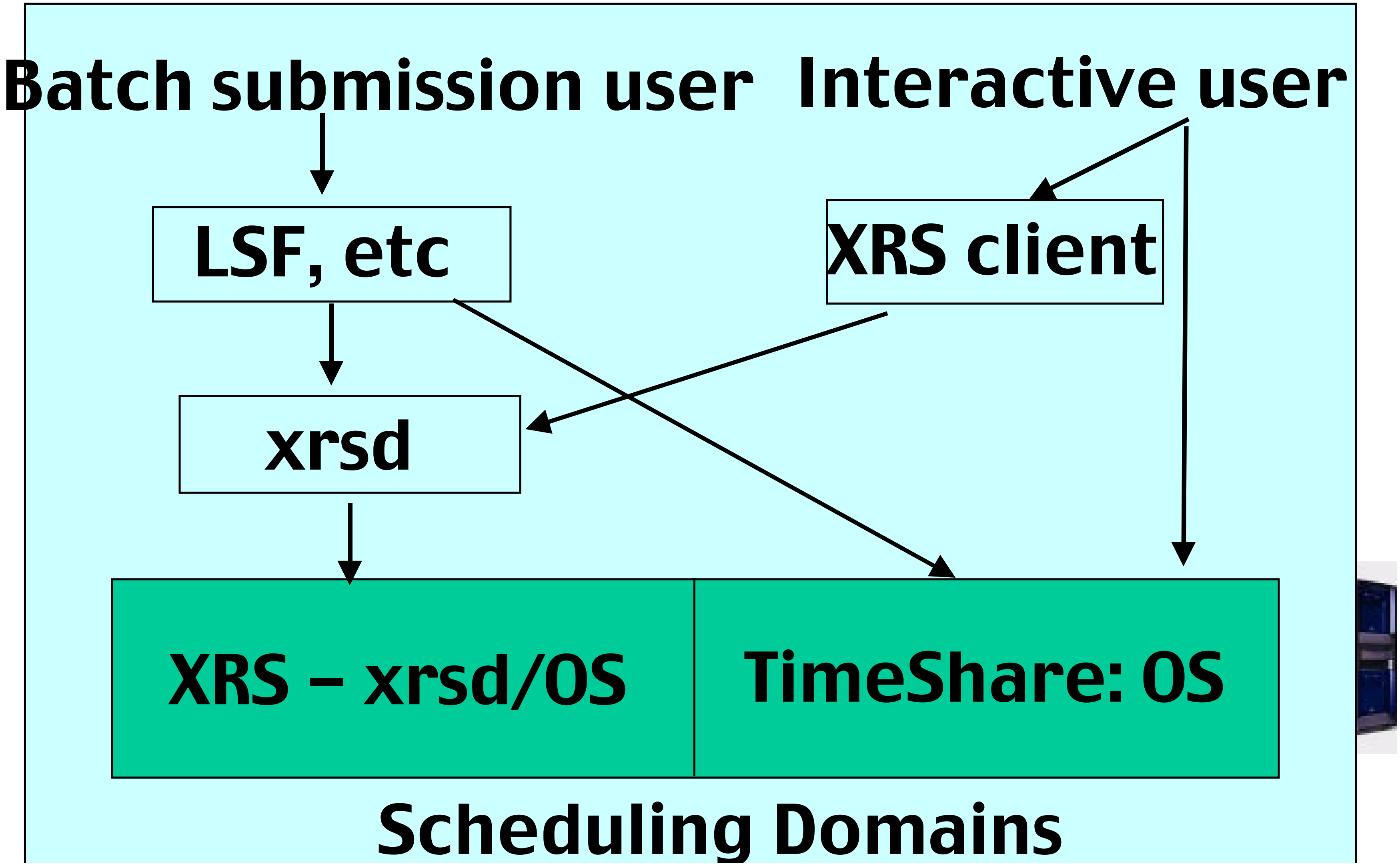
Next Generation Resource Scheduler

- Manages the allocation of resources for jobs
  - Guaranteed resource reservations
- Flexible resource reservation framework
  - Customer extensible to meet unique scheduling requirements
  - User specific placements
- Published API

IRIX Extensible Resource Scheduler (XRS)

sgi

XRS Scheduling Domains



IRIX Extensible Resource Scheduler (XRS)

Scheduling Partitions

- The XRS scheduling domain can be organized into various scheduling partitions
- A scheduling partition is a collection of resources and the scheduling policy that manages those resources

IRIX Extensible Resource Scheduler (XRS)

Resources to be managed initially:

- CPU – speed, cache size and speed, local memory size, neighbor cpus
- Memory – allocations managed per-node, cross referenced against resident cpus
- Topology – user can provide dplace-compliant placement file

XRS Scheduling Policies

sgi

- Predictive
  - predictive completion times, no preemption
- Availability
  - like predictive with repack if jobs complete early
- Priority
  - like availability with priority scheme and re-ordering
- Shared
  - allows over-subscription of renewable resources
- Preemptive
  - user may preempt running job. Running job is suspended , or checkpointed. Supplementary to all but Predictive.

IRIX Extensible Resource Scheduler (XRS)

sgi

Status

- Requirements and Concept documents are complete
- Research, prototyping, and design in progress
- Beta testing in Q2CY00 at Boeing
- General availability planned for IRIX 6.5.9 (Q3CY00)
- Integrating IRIX XRS with LSF.

Workload Management

sgi

Partnership with platform computing

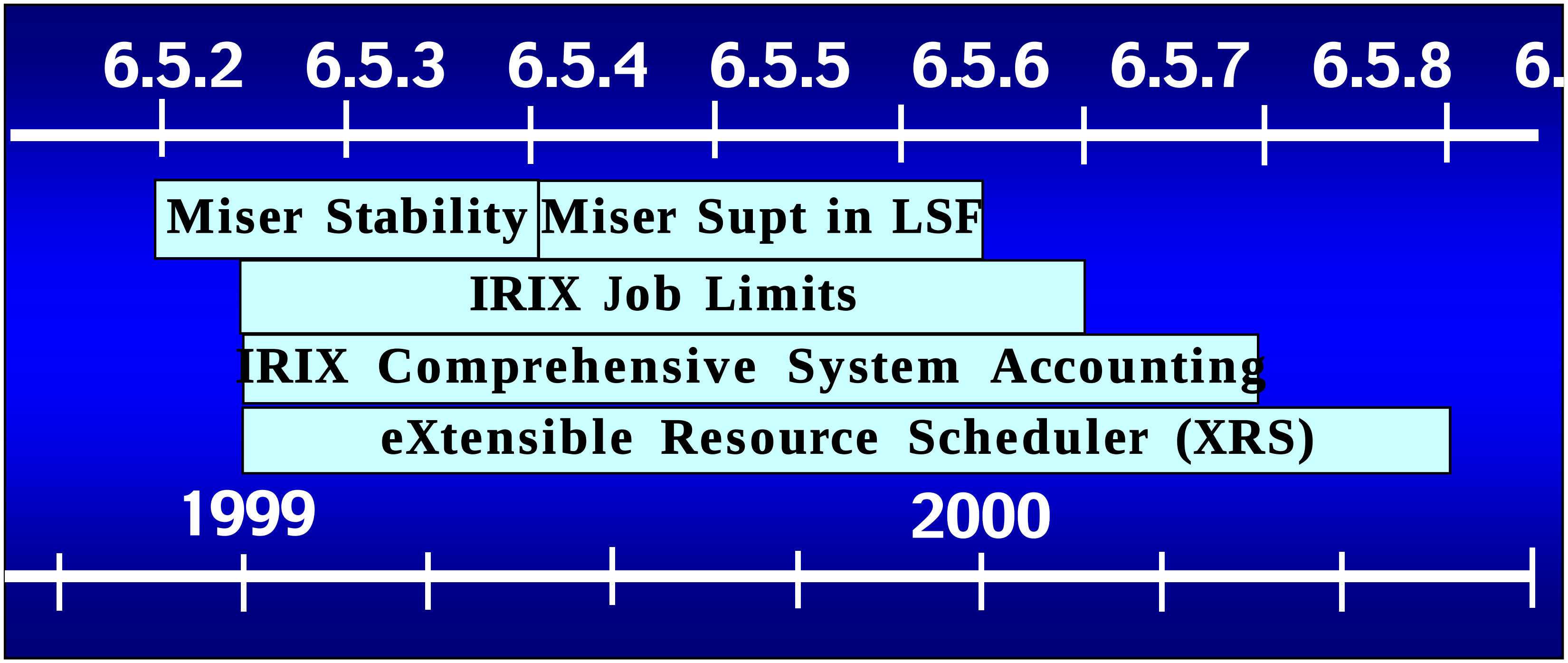
- LSF 3.2 for IRIX, UNICOS & UNICOS/mk available now
- LSF will support SNx & SVx
- MPT supported with LSF Parallel available now
- NQE features in LSF 4.0 available in Q4CY99:
  - File Transfer Agent (FTA)
  - Improved output file handling
  - UNICOS accounting support
  - Job-based limits for major resources
- Integrating IRIX job limits, CSA, Miser, and XRS with LSF

Workload Management

Network queuing environment (NQE)

- NQE feature development is complete for SGI platforms with NQE 3.3
- NQE support for SGI platforms (including SV1) continues through 2004
- NQE is retired for non-sgi platforms

IRIX Resource Management Roadmap



6.5.2 6.5.3 6.5.4 6.5.5 6.5.6 6.5.7 6.5.8 6.

Miser Stability

Miser Supt in LSF

IRIX Job Limits

IRIX Comprehensive System Accounting

eXtensible Resource Scheduler (XRS)

1999

2000

Summary

- IRIX Job Limits in IRIX 6.5.7 (Q1CY00)
- IRIX CSA in IRIX 6.5.8 (Q2CY00)
- Miser much more reliable and performs better in IRIX 6.5.4
- IRIX XRS in IRIX 6.5.9 (Q3CY00)
- LSF is our workload management solution
- NQE 3.3 supported on SGI platforms through 2004
- NQE retired on non SGI platforms