Irix: Origin Software Roadmap

Update of System Management Software for Large Origin Systems

Daryl Coulthart
Silicon Graphics Inc
dbc@cray.com

SiliconGraphics
Computer Systems

System management for large Origin systems

- **Provide Data Center quality and HPC functionality**
  - **Make progressive improvements in Irix**
  - **Move selected UNICOS capabilities to Irix**
  - **Requirements from Resource management forum (Feb 98)**
- **Components**
  - OS scaling
  - Kernel Scheduling
  - Batch queuing
  - Message passing
  - Checkpoint / Restart
  - Cluster Infrastructure
  - Resource Limits
  - Accounting

OS Scaling

- **Irix 6.4** **32 PE support - 4Q1996**
- **Irix 6.4 update** **64 PE support - 3Q1997**
- **Irix 6.5** **128 PE support - 3Q1998**
- **Cellular Irix** **256 PE support - SN1 timeframe**

Kernel Scheduling

- **Share II**
  - **fairshare**
  - **Irix 6.5**
- **Miser**
  - **schedule CPUs, memory and cpu time**
  - **Irix 6.5**
- **Miser API**
  - **Cluster scheduling**

Batch Queuing

- **NQE 3.1**
  - **Initial Origin support**
  - **Released 4Q1996**
- **NQE 3.2**
  - **Irix 6.4 support**
  - **64 bit version, 64 bit process limits, cpr support**
  - **Released 1Q1997**
- **NQE 3.2.2**
  - **soft job limits**
  - **Released 3Q1997**
- **NQE 3.3**
  - **DCE ticket forwarding**
  - **Miser**
  - **Irix 6.5**
  - **Released 2Q1998**

64 bit soft limits Miser sched

Message Passing

- **MPI 1.2**
  - **64 PE support**
  - **LSF support**
  - **Released 1Q1998**
- **MPI 1.2.1**
  - **Miser support**
  - **16 X 128 PE support**
  - **single executable/single system CPR**
  - **Release 2Q1998**
- **MPI 1.3**
  - **48 X 128 PE support**
  - **Planned 4Q1998**
- **MPI X.Y**
  - **Cluster CPR Phase 1**
  - **Planned 4Q1998**

PE scaling cpr 48x128 cluster cpr

Checkpoint / Restart

- **CPR 1.0 - Irix 6.4**
  - **Interactive, libmp, basic batch**
  - **Released 4Q1996**
- **NQE 3.2**
  - **Batch integration**
  - **Released 1Q1997**
- **CPR 1.1 - Irix 6.4 update**
  - **Pthreads, fixes**
  - **Released 3Q1997**

Checkpoint / Restart (cont.)

- **MPI 1.2.1/Array 3.1**
  - **Mpi single machine checkpoint**
  - **Released 2Q1998**
- **CPR 1.2 - Irix 6.5**
  - **Fetchop, shmem**
  - **Released 3Q1998**
- **MPI cluster cpr**
  - **MPI 1.X + Array 3.3**
  - **Planned 4Q1998**

Cluster Infrastructure

- **Array 3.0**
  - **Irix 6.4**
  - **8 x 32 node support**
  - **4Q1996**
- **Array 3.1**
  - **Irix 6.5**
  - **MPI single machine cpr**
  - **16 x 128 node support**
  - **single machine miser**
  - **2Q1998**

8x32 16x128

Cluster Infrastructure (cont.)

- **Array 3.2**
  - **Source merge**
  - **UNICOS commands**
  - **3Q1998**
- **Array 3.3**
  - **48x128 support x 128 node support**
  - **4Q1998**

8x32 16x128 UNICOS 48x128

Resource Limits

- **Irix 6.4**
  - **64 bit limits**
  - **4Q1996**
- **NQE 3.2**
  - **process limits**
  - **1Q1997**
- **NQE 3.2.2**
  - **soft job limits**
  - **monitor/sample resources**
  - **4Q1997**

process softjob

Resource Limits(cont)

- **Interactive User limits**
  - **Similar to NQE softjob limits**
  - **Interactive job is sum of all interactive processes**
  - **Uses early version of UDB for administration**
  - **Limits (cpu time, memory, #PE~=# processes**
  - **Support will be dropped after UDB/Job limits implemented**
  - **Available 1Q1998 as a web release**
- **UDB/JOB/Kernel**
  - **Job id**
  - **New job number created at login, batch job, miser session**
  - **Available 1Q1999 in an Irix release**

process softjob/iaud kernel/udb

Accounting

- **Irix process accounting**
- **NQE accounting text file**
- **ASH Irix extended accounting/Perfacct**
- **Cray style extended accounting**
  - **Job level accounting**
  - **Support project id**
  - **Cray style accounting reports**
  - **Available 1Q1999 in an Irix release**

Summary

OS	32PE	64PE	128PE	256PE
Scheduling		Share II	Miser	Cluster api
Batch	64 bit	soft limits		Miser sched
MPI	PE scaling		48x128	cluster cpr
CPR	pthreads	MPI		cluster
Array	8x32	UNICOS	16x128	48x128
Limits	process	softjob/iaud		kernel/udb
Accounting	process	ASH		Craylike