Cray User Group Conference
June 15–19, 1998 Stuttgart, Germany

Partitioning the Origin 2000(TM)

Allan Christie
Member Technical Staff
Silicon Graphics Computer Systems
ajc@sgi.com

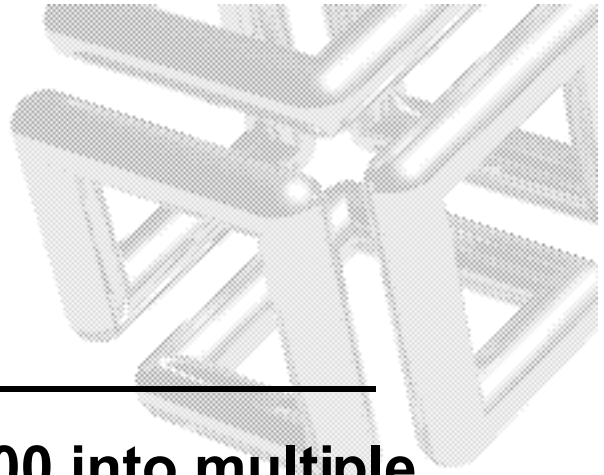
SiliconGraphics Computer Systems

IRIX(TM) 6.5 Partitions on the Silicon Graphics Origin 2000

- **Overview**
  - What is a Partition, and what can I do with it
- **Configurations**
  - Supported hardware configurations
- **Components**
  - What is required to use partitioning
  - What roles do the different components play
- **Customer Visible Features**
  - Interfaces provided to the user and system administrator
  - Introduction to setting up partitions

SiliconGraphics Computer Systems

What is a Partition?



- **A method to divide a single Origin 2000 into multiple distinct systems**
  - Not a way to connect multiple Origin 2000 systems with Craylink
  - A way to divide an Origin 2000 into multiple IRIX systems
- **Foundation for Cellular IRIX**
  - A CELL is a partition

SiliconGraphics Computer Systems

What is a Partition?

Continued ...

- **Each Partition is independent**
  - Runs separate copy of IRIX
    - Has own I/O
    - Has dedicated boot and swap device(s)
    - Has dedicated Console
  - May be booted, power cycled, loaded with software independently
  - Partition's memory and Craylink bandwidth NOT affected by other partitions

What is a Partition?

Continued ...

- **Hardware supported partition isolation**
  - Memory protections
  - I/O device protections
  - CPU protections
  - Reset Propagation
  - Block Transfer Engine
- **Hardware support dictates granularity**
  - Hardware requires partitions are a multiple of Modules

What is a Partition

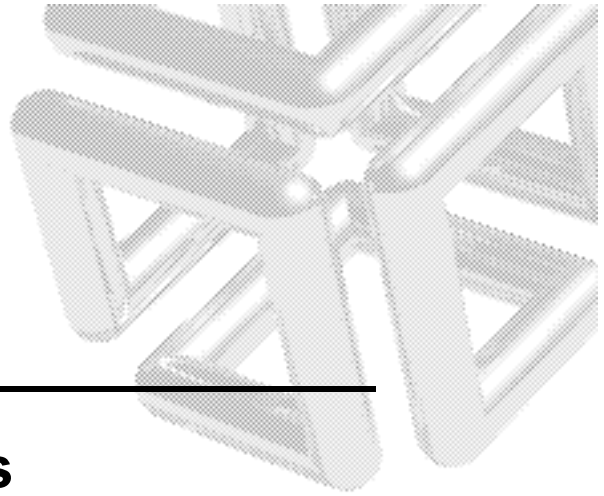
Continued ...

- **Rack system can be**

  2 x 1 module partitions
  - Requires I/O board in upper and lower module
- **Desk-side system can be connected with others to form 2 or more partitions**
- **Origin 200(TM) can not run as part of a partitioned system**

SiliconGraphics Computer Systems

Configurations



- **Requirements for valid configurations**
  - All partitions are self contained. No intra-partition Craylink traffic can travel outside the partition
  - All partitions are fully interconnected. Inter-partition Craylink traffic must travel within the source and destination partition only
  - A Partition is comprised of Modules
- **Version 5 of the HUB is required on all nodes**
  - hinv -v

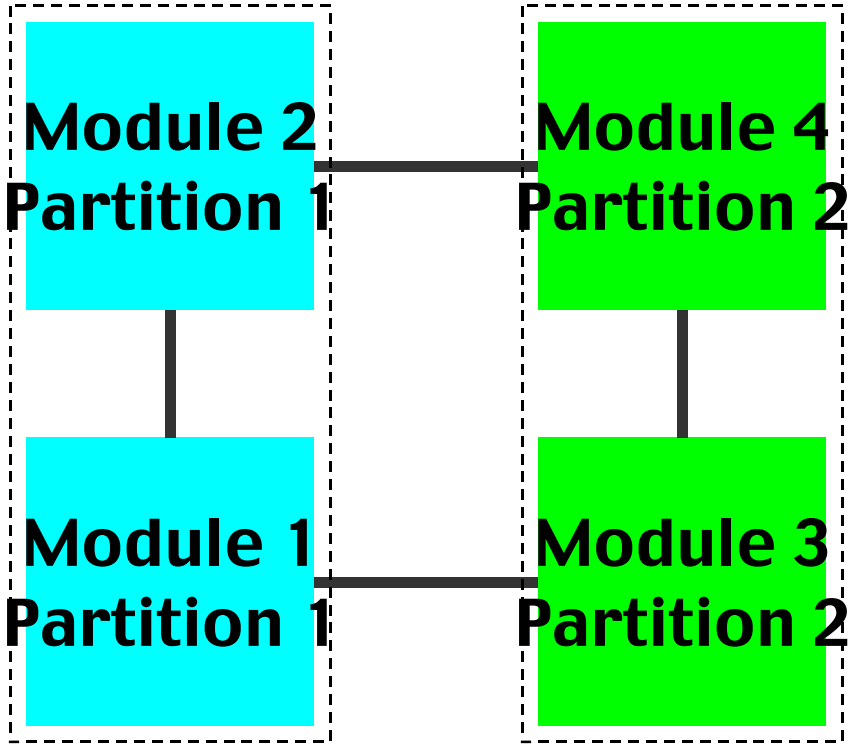
    > HUB in Module X/Slot Y: Revision 5 (enabled)

SiliconGraphics Computer Systems

Configurations

Continued ...

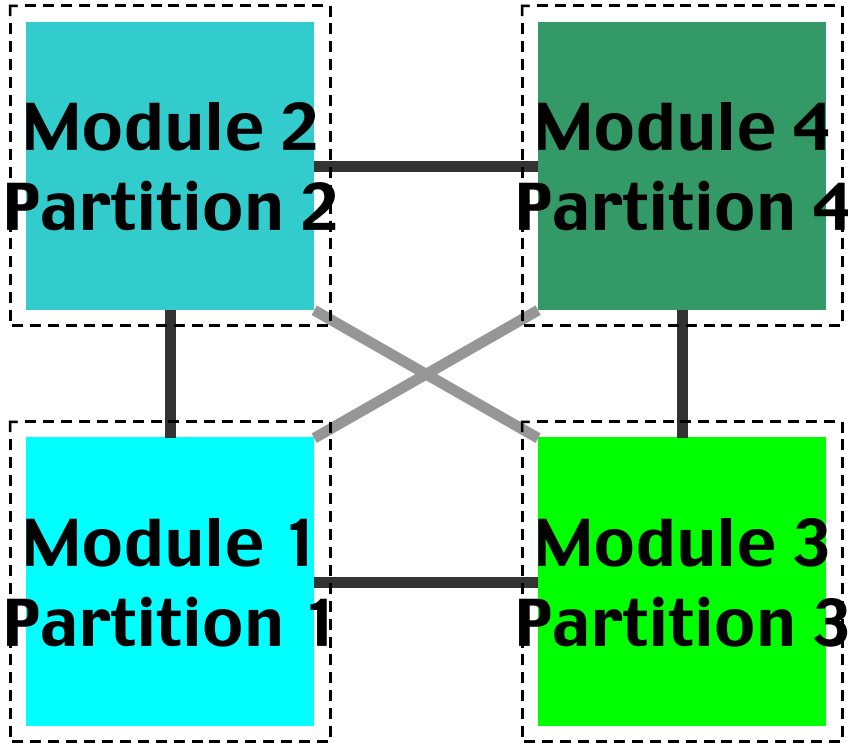
4 modules - 2 partitions



Configurations

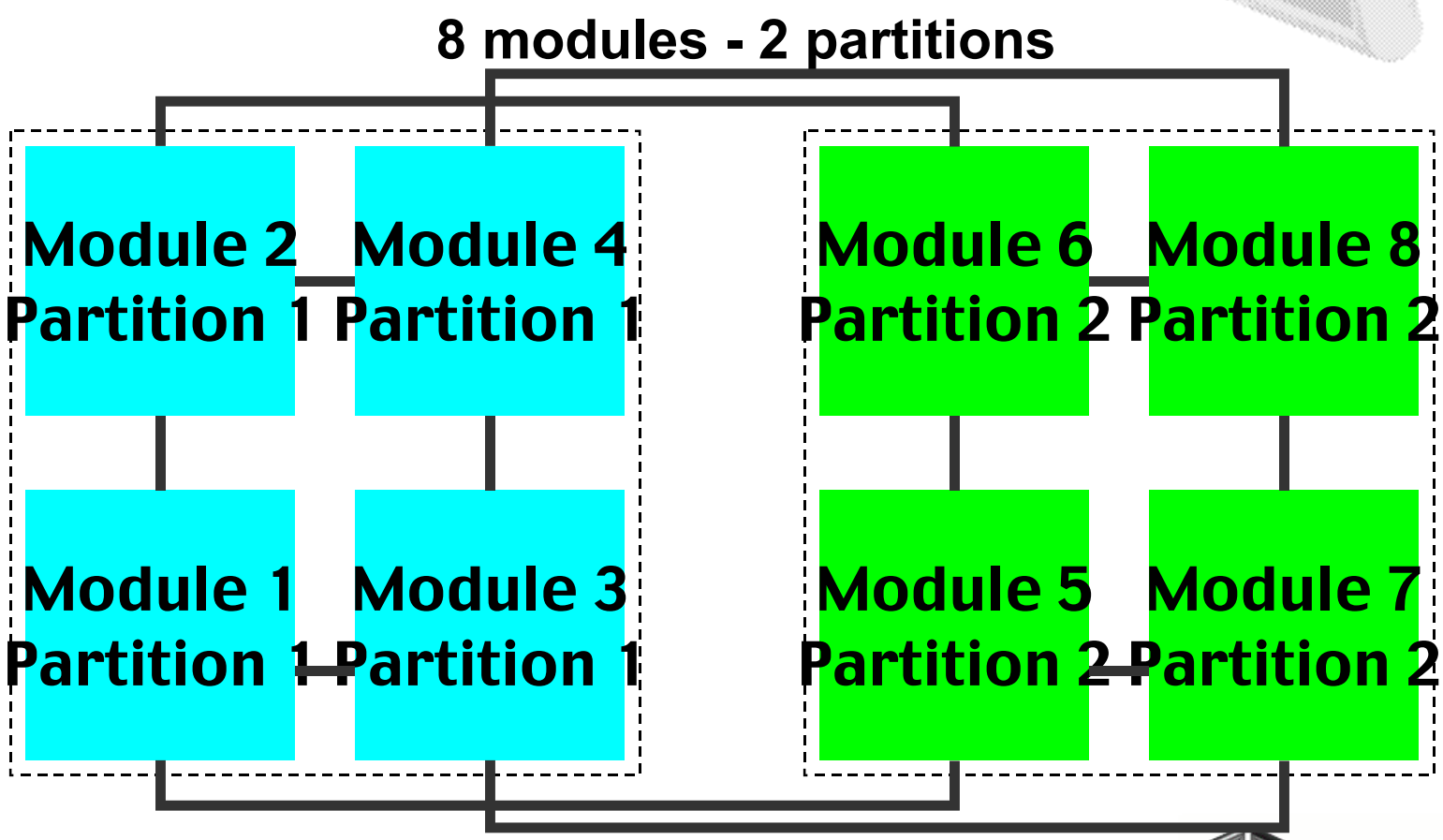
Continued ...

4 modules - 4 partitions



Configurations

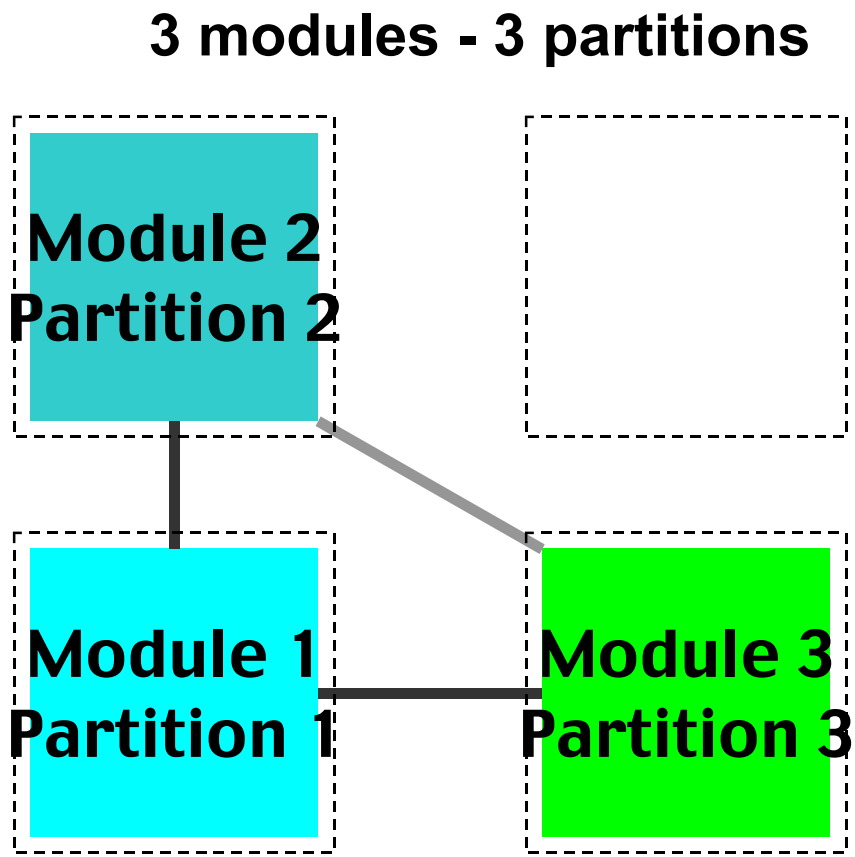
Continued ...



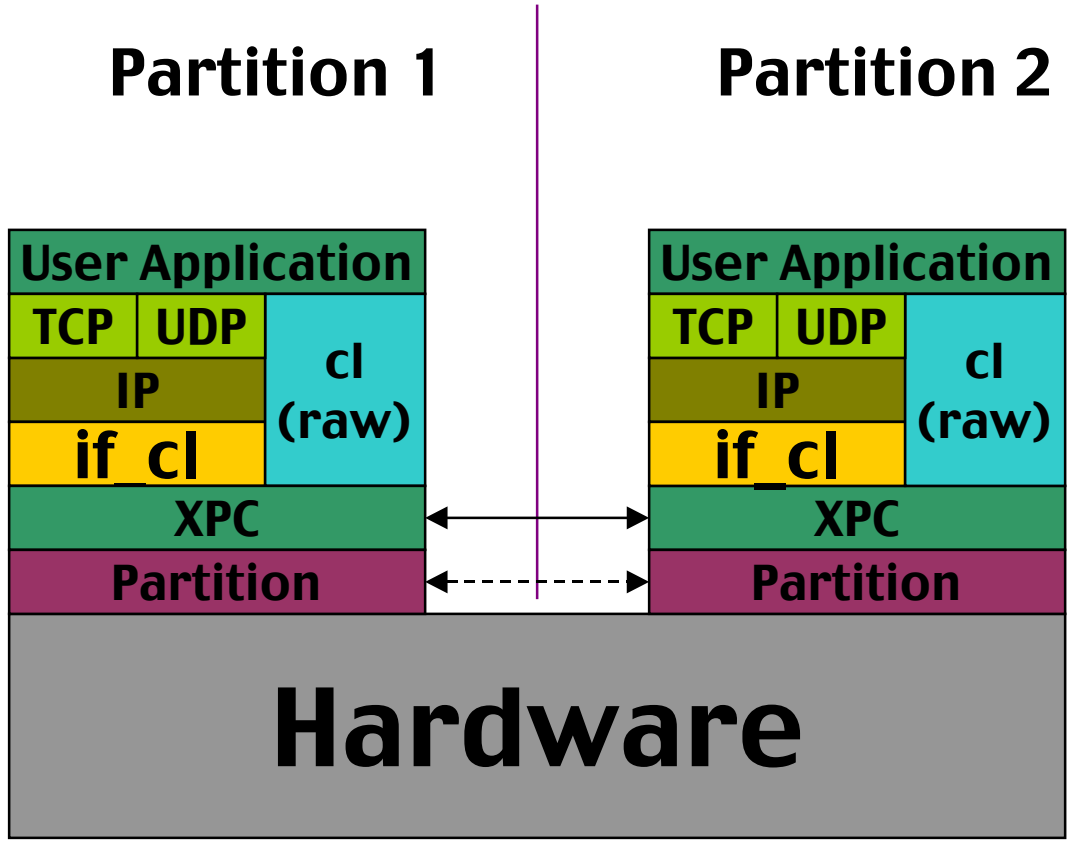
SiliconGraphics
Computer Systems

Configurations

Continued ...



Software Components



SiliconGraphics Computer Systems

Software Components

Continued ...

- **Partition Module**
  - Protects memory and CPU from other partitions
  - Recognizes the appearance and disappearance of remote partitions
- **XPC Module**
  - Reliable message passing between partitions
  - Required for all cross partition communication (if_cl and cl)

Software Components

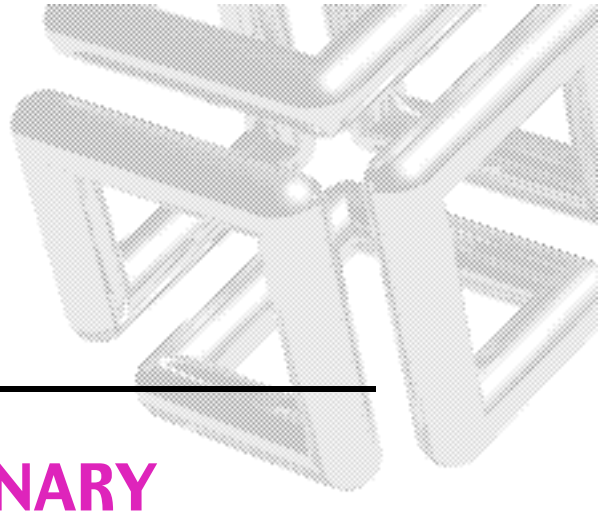
Continued ...

- **lf_cl network driver**
  - Provides standard network interface semantics
  - Allows TCP/IP UDP/IP between partitions
    - NFS and BDS
    - sockets
  - Requires XPC module for actual data transfer
- **cl driver**
  - Provides raw device byte-stream semantics
  - read/write raw devices in 2 partitions to transfer data
  - Requires XPC module for actual data transfer

Customer Visible Features

- **mkpart command**
  - Configure or query partition information
- **mkpartd**
  - Daemon which communicates with daemons on other known partitions
- **cl0 network interface**
  - IP based network interface
- **/hw/xplink/raw/<partition>/<device> raw devices**
  - read/write/select byte stream interface

TCP/IP Performance



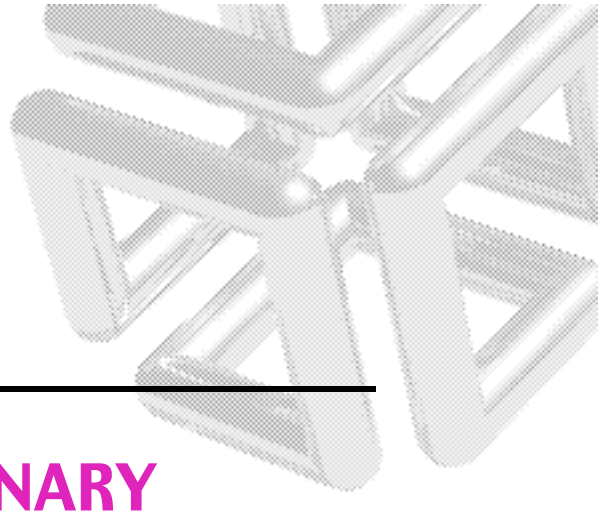
Performance numbers PRELIMINARY (Pre-release IRIX 6.5)

Socket Buffer Size	Transmit Rate	Receive Rate	% data dropped
1048576	200+ MB/s	200+ MB/s	0 %
524288	200+ MB/s	200+ MB/s	0 %
262144	170+ MB/s	170+ MB/s	0 %
131072	95+ MB/s	95+ MB/s	0 %
65536	55+ MB/s	55+ MB/s	0 %

2-module, 2 partition system (8 CPUs per partition), 4MB SCACHE, 195MHz

SiliconGraphics Computer Systems

UDP/IP Performance



Performance numbers PRELIMINARY (Pre-release IRIX 6.5)

Socket Buffer Size	Transmit Rate	Receive Rate	% data dropped
1048576	270+ MB/s	270+ MB/s	< 1%
524288	266+ MB/s	263+ MB/s	< 1%
262144	270+ MB/s	255+ MB/s	5.5 %
131072	275+ MB/s	220+ MB/s	20 %
65536	275+ MB/s	160+ MB/s	45 %

2-module, 2 partition system (8 CPUs per partition), 4MB SCACHE, 195MHz

SiliconGraphics Computer Systems