# Cray's Scalable I/O System: The GigaRing™ System Area Network

*R. Kent Koeninger*, Cray Research, Inc., 655F Lone Oak Drive, Eagan, Minnesota 55121

**ABSTRACT:** *This paper describes the new properties of the GigaRing system and what it means to Cray sites converting to GigaRing. It introduces the concept of System Area Networks with specific examples from GigaRing technology.*

## 1 Introduction

GigaRing is a ring-based I/O system that connects Cray (and other) hosts with I/O nodes for disks, tapes, and networks. It can also be used as a very fast "System Area Network" (SAN). GigaRing technology lowers the cost of I/O significantly (when compared with Model-E based systems), while increasing bandwidths and improving the flexibility of upgrades.

## 2 Basic GigaRing Architecture

In 1996, GigaRing will connect to CRAY T3E, CRAY T90, and CRAY J90se hosts. It provides multiple GigaRing channels per host with over a gigabyte-per-second of bandwidth per channel. The I/O nodes supported include:

- Disks (SCSI, Fibre Channel, IPI)
- Tapes (SCSI, ESCON, Block Mux)
- Networks (ATM, HIPPI, FDDI, Ethernet)
- CRAY SSD-T90

The GigaRing channels are fault tolerant. They use two counter-rotating rings that allow a node or a ring to fail while maintaining connectivity among the non-failing components.

GigaRing channels are based on IEEE Scalable Coherent Interface (SCI) technology with enhancements. (It is not strictly compliant with the SCI standard.)

## 3 System Area Networks

System Area Networks (SANs) have higher bandwidths and lower latencies than Local Area Networks (LANs) or Wide Area Networks (WANs). To achieve this superior performance, SANs are less scalable than LANs or WANs and typically connect a few or a few dozen nodes. SANs are typically used within machine cabinets or machine rooms.



Example SANs:

- SCI Technology: GigaRing, Dolphin
- Others: HIPPI, Myrinet, HIPPI 6400, Fibre Channel

GigaRing is an open SAN. The specification is open and the technology can be (and has been) licensed by other companies for a variety of SAN uses.

Cray expects a convergence in the market on a dominant SAN standard. Today, most computer companies support SANs, but with little agreement on how to interoperate across the various SAN technologies. Cray will work with industry players to help form a SAN standard and will migrate GigaRing to this new standard.

GigaRing enhancements to the SCI standard include:

- Direct Memory Access (DMA) for low overhead transfers.
- Distance extension: 200 meters between nodes.
- High bandwidth: 2 x 800 megabytes per second.
- Resiliency: Counter rotating rings with folding and masking.
- Security: Membranes across client interfaces to filter packets and protect systems.
- Large buffers for large bandwidths.
- Flow control: Multiple virtual channels and end-to-end adaptive congestion control mechanisms.
- Integrated error reporting for monitoring error conditions on channels and nodes.

## 4 GigaRing Bandwidths

As a general rule, the GigaRing channels provide more bandwidth than needed to drive peripherals. Measured in macho-bytes, each channel provides over a gigabyte-per-second of maximum payload.

The GigaRing client chip interfaces are faster than any individual planned node. Example client bandwidths:

- CRAY T3E: 1 to 128 channels
  - E.g., ~100 MB/s per air-cooled PE = ~6 gigabytes/sec for 64 PEs
  - E.g., ~50 MB/s per liquid cooled PE = ~12 gigabytes for 256 PEs
- CRAY T90: 1 to 32 channels
  - E.g., 8 channels at ~600 MB/s each = ~5 gigabytes/sec
- CRAY J90se: 1 to 4 channels
  - E.g., 4 channels at ~200 MB/s each = ~800 megabytes/sec
- Disk, tape and network nodes:
  - ~100 to ~200 MB/s each
  - Excess host bandwidth per channel to drive peripherals

(All example bandwidths are quoted in one direction. Bidirectional (full-duplex) total bandwidths are faster.)

## 5 Comparisons with Model-E I/O

Model-E I/O Subsystems provide the fastest I/O to date. They cost more than GigaRing I/O with major cost in the frames, IOCs, and controllers. The IPI disks popular with Model-E systems are significantly more expensive per-byte than Fibre Channel or SCSI based disks. Model-E systems are less scalable than GigaRing systems and are generally limited to a handful of 200 MB/s HISP channels.

Commodity-technology disks (such as DD-308 Fiber Channel disks) dominate the cost of GigaRing I/O. These disks deliver ~1 MB/s of bandwidth for each gigabyte of storage; for hundreds of megabytes-per-second of bandwidth, one needs hundreds of gigabytes of disks. For example, to drive the 12 gigabytes of bandwidth on a 256 PE CRAY T3E would require 12 terabytes of disk. The Fiber Channel disks based on commodity-disk technology cost less than the IPI disks, but 12 terabytes is beyond the disk budgets of most sites.

## 6 Product Status

GigaRing hardware is working at Cray now. CRAY T3E systems read and write SCSI devices across GigaRing channels through MPN-1 nodes. In fact, Cray shipped a working CRAY T3E system to Pittsburgh Supercomputer Center, the week after the Barcelona CUG.

Cray will roll out mainframe and peripheral nodes in stages throughout 1996. The CRAY T3E systems with MPN-1 and HPN-1 nodes will ship initially with CRAY T90, CRAY J90se and other peripheral nodes to follow later in 1996.

## 7 CRAY SSD-T90

The CRAY SSD-T90 is a solid-state-memory device for GigaRing based CRAY T90 systems. It provides high-bandwidth and low latency for the UNICOS I/O buffering techniques, such as ldcache, SDS, and SSD file systems. It supports back-door I/O between the SSD and disks. A typical CRAY SSD-T90 peripheral will deliver over a gigabyte-per-second of bandwidth for large transfers.

## 8 GigaRing Features in 1996

The GigaRing hardware will ship in stages throughout 1996:

- SCSI disks and tapes
- Ethernet, FDDI, ATM OC-3, HIPPI networks
- IPI, Fibre Channel, & ND disks
  - RAID-5: ND (HIPPI) disks
  - RAID-3: Fibre Channel & ND disks
- Block Mux and ESCON tapes

In 1996, Cray will support a single host per GigaRing channel. TCP/IP communication among hosts will work through standard network interfaces such as HIPPI, FDDI, ATM, or Ethernet.

In 1996, Cray will support alternate paths for primary and secondary resilient paths to Fibre Channel and IPI disks.

## 9 GigaRing Features in 1997

In 1997, Cray will ship GigaRing nodes with ATM OC-12 support (77 MB/s). GigaRing software for RAID-5 support on Fibre Channel disks is also slated for 1997.

In 1997, Cray will support direct host-to-host communication over GigaRing channels. This will increase the speed of TCP/IP and UDP/IP communication while reducing the communication hardware necessary for host-to-host communication. The IP stack will be implemented with GigaRing DMAs. NFS, NFS V3, DFS, PVM, and other IP based protocols will benefit from this increased IP performance.

Cray is studying faster-than-IP GigaRing transport mechanism that may result in substantially faster NFS transport rates than seen to date.

SFS is supported on UNICOS and UNICOS/mk hosts across GigaRing to Fibre Channel disks. Cray expects most sites to prefer NFS or DFS distributed-file solutions instead of SFS, especially as the speed of NFS and DFS increases on GigaRing.

## 10 Summary

The GigaRing technology is a high-end System Area Network solution with room to grow for I/O needs far into the future. The GigaRing channel and host-interface bandwidths will rarely be a limiting factor in I/O configurations. GigaRing commodity technology improves the cost of I/O significantly, when compared with Model-E based systems.