GSN on SGI’s new NUMAFlex Architecture - Origin 3000

Anthony F. Voellm

High Performance Network Engineering

voellm@sgi.com

Benchmarks and input provided by Chad Carlin and David Powell.

Presentation Overview

-*Definitions*

-*Origin 3000 Architecture (NUMAFlex)*

-*Comparison of O3000 vs. O2000*

-*Current news*

-*Review*

-*Sources for more information*

Definitions

ST - Scheduled Transfer

STP - Scheduled Transfer Protocol

GSN - Gigabyte System Network

Origin 3000 Architecture

Overview

- *Origin 3000 is SGI’s NEW follow-on to the Origin 2000*

- *The architecture is called NUMAFlex because of the highly configurable design*

- *It is double the bandwidth of Origin 2000.*

- *Scalable to 1024 processors*

- *Fault tolerant and redundant*

Origin 3000 Architecture

Bricks

- *Designed as a series of bricks*
  - C-brick - Hold 0,2 or 4 R12K+ Processors
  - R-brick - Router Interconnect (NumaLink3)
  - I-brick - Base I/O Module (Con. TTY, …)
  - P-brick - PCI expansion (PCIX next)
  - X-brick - XIO expansion (used by GSN)
  - G-brick - Graphics expansion
  - D-brick - JBOD Disk storage
  - Power Bay for N+1 redundant Power

Origin 3000 Architecture

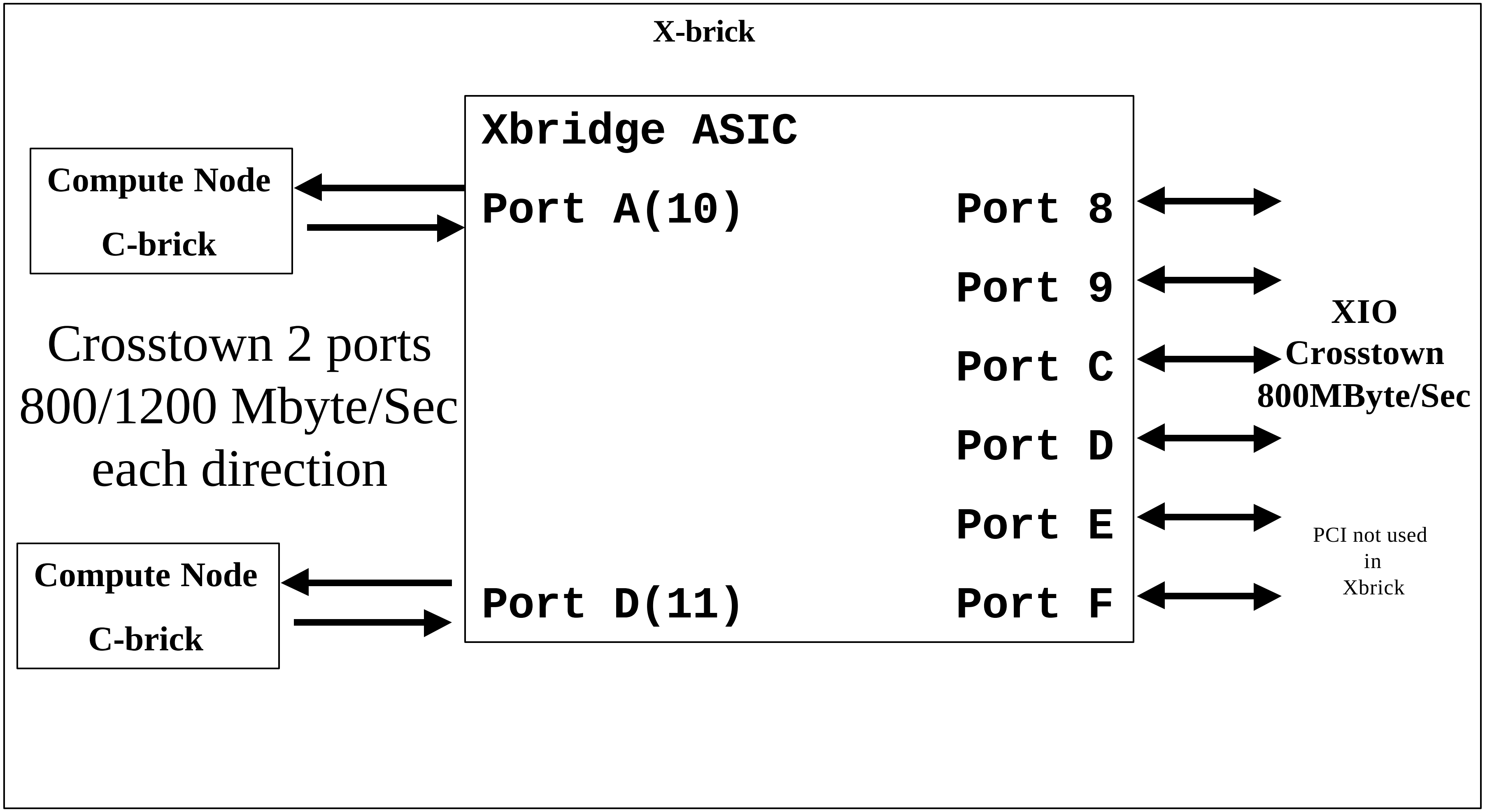
Explanation of Bricks used by GSN

-*The following bricks are used by GSN*

- X-brick
  - Used by the GSN NIC
- C-brick
  - The C-brick handles GSN interrupts and processes IP and STP traffic
- R-brick
  - GSN data can travel though the router

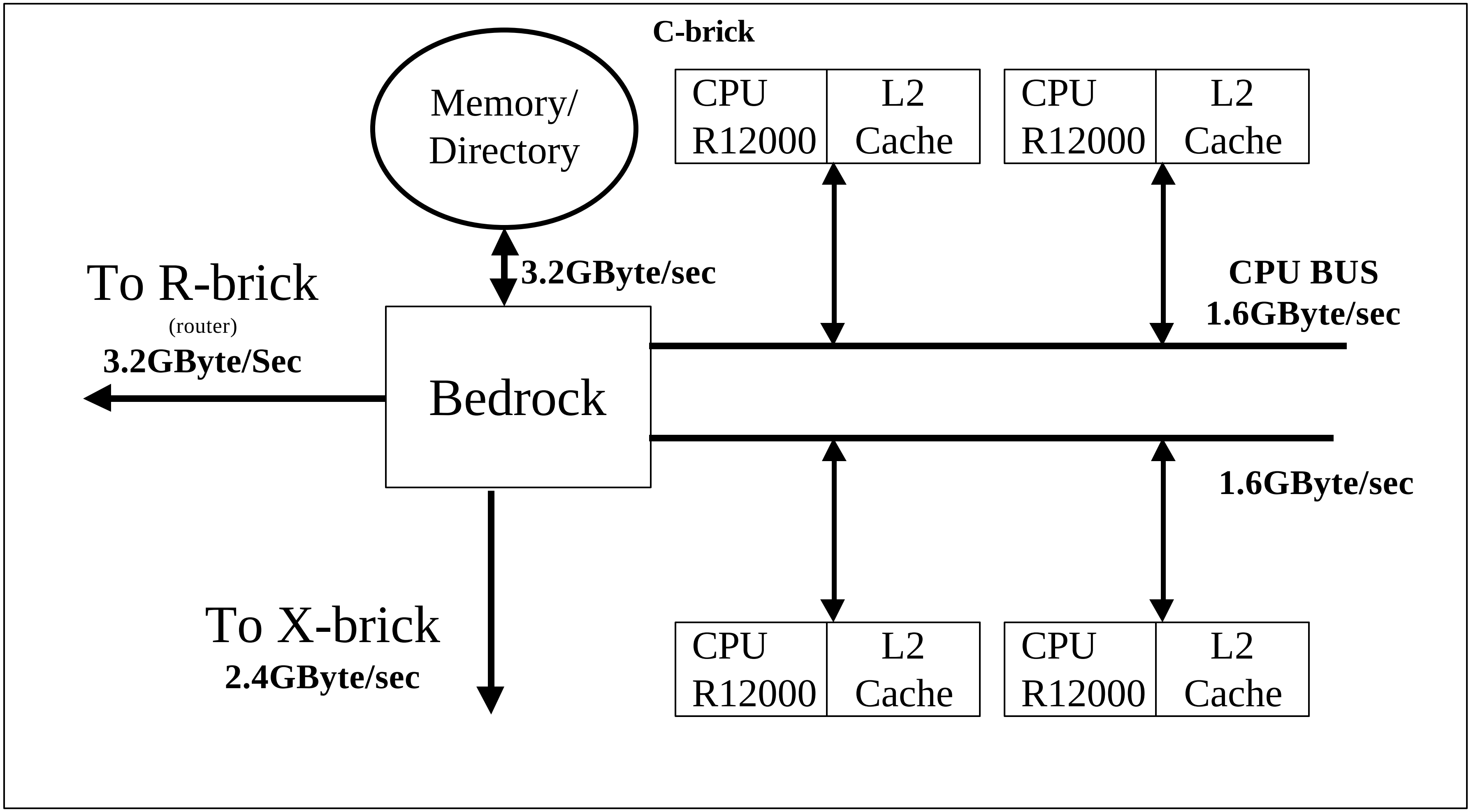
Origin 3000 Architecture

Explanation of Bricks used by GSN: X-brick



Origin 3000 Architecture

Explanation of Bricks used by GSN: C-brick



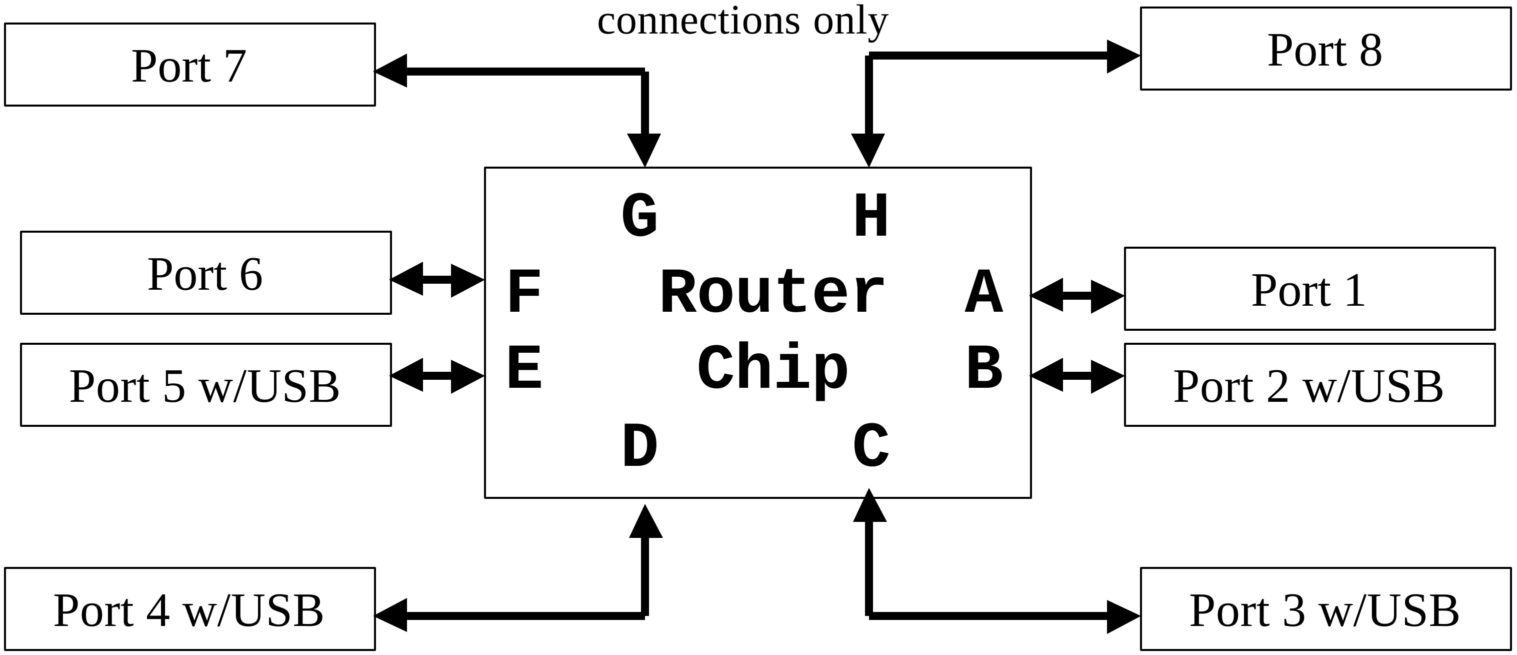
All Bandwidth numbers are full-duplex

Origin 3000 Architecture

Explanation of Bricks used by GSN: R-brick

R-brick

Port 1,6,7 & 8
are R-brick to R-brick
connections only



Port 7

Port 8

Port 6

Port 5 w/USB

G H
F Router A
E Chip B
D C

Port 1

Port 2 w/USB

Port 4 w/USB

Port 3 w/USB

Port 2,3,4,5
are R-brick to C-brick
or
R-brick to R-brick

Origin 3000 Architecture

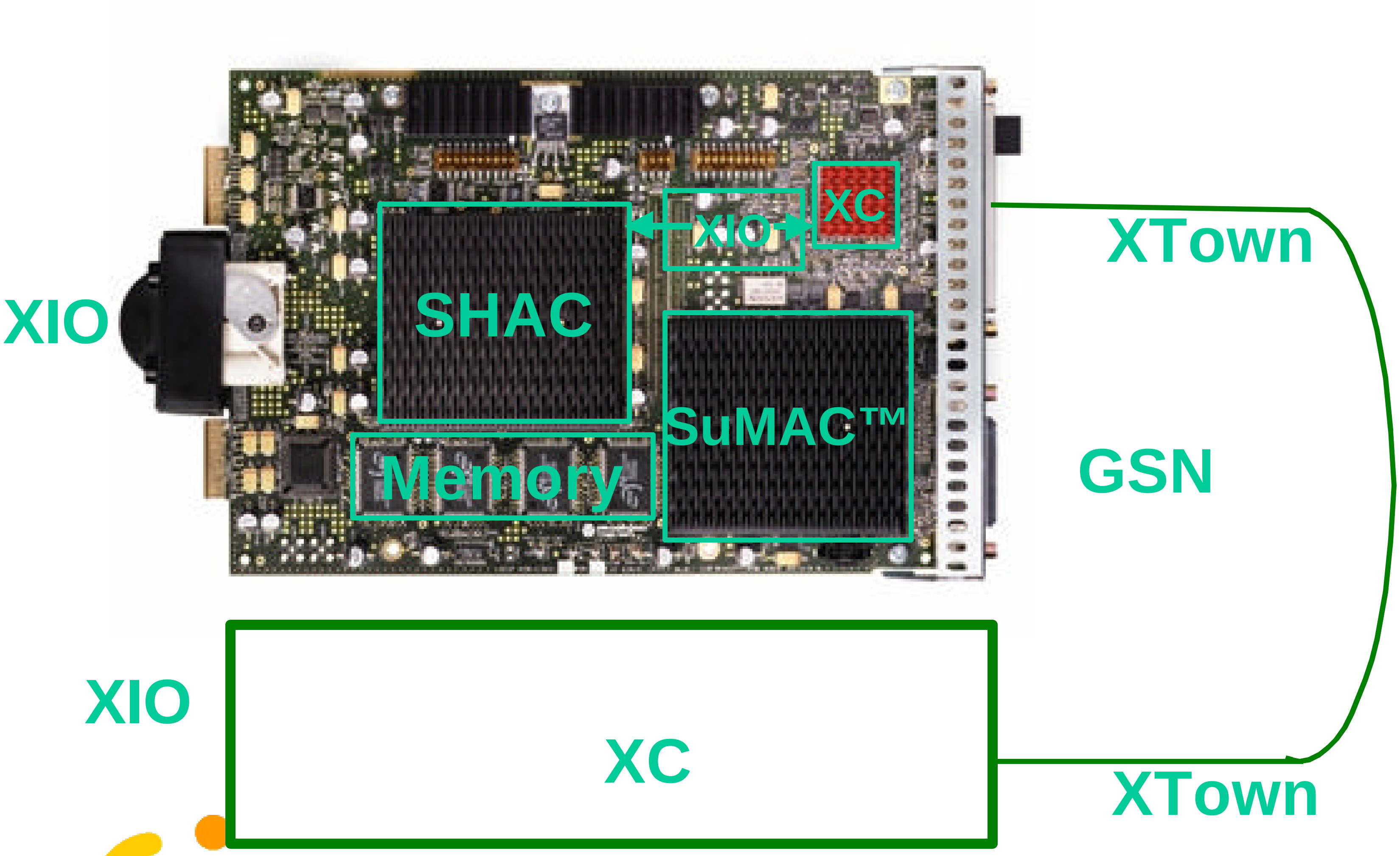
GSN

-*GSN is a two board XIO set*

- XT0 (the primary) is where the SUMAC and SHAC are located. These are the brains of the SGI GSN implementation.
- XT1 (the secondary) is an XTOWN board used to improve bandwidth by providing a shorter path to Origin memories located near it.

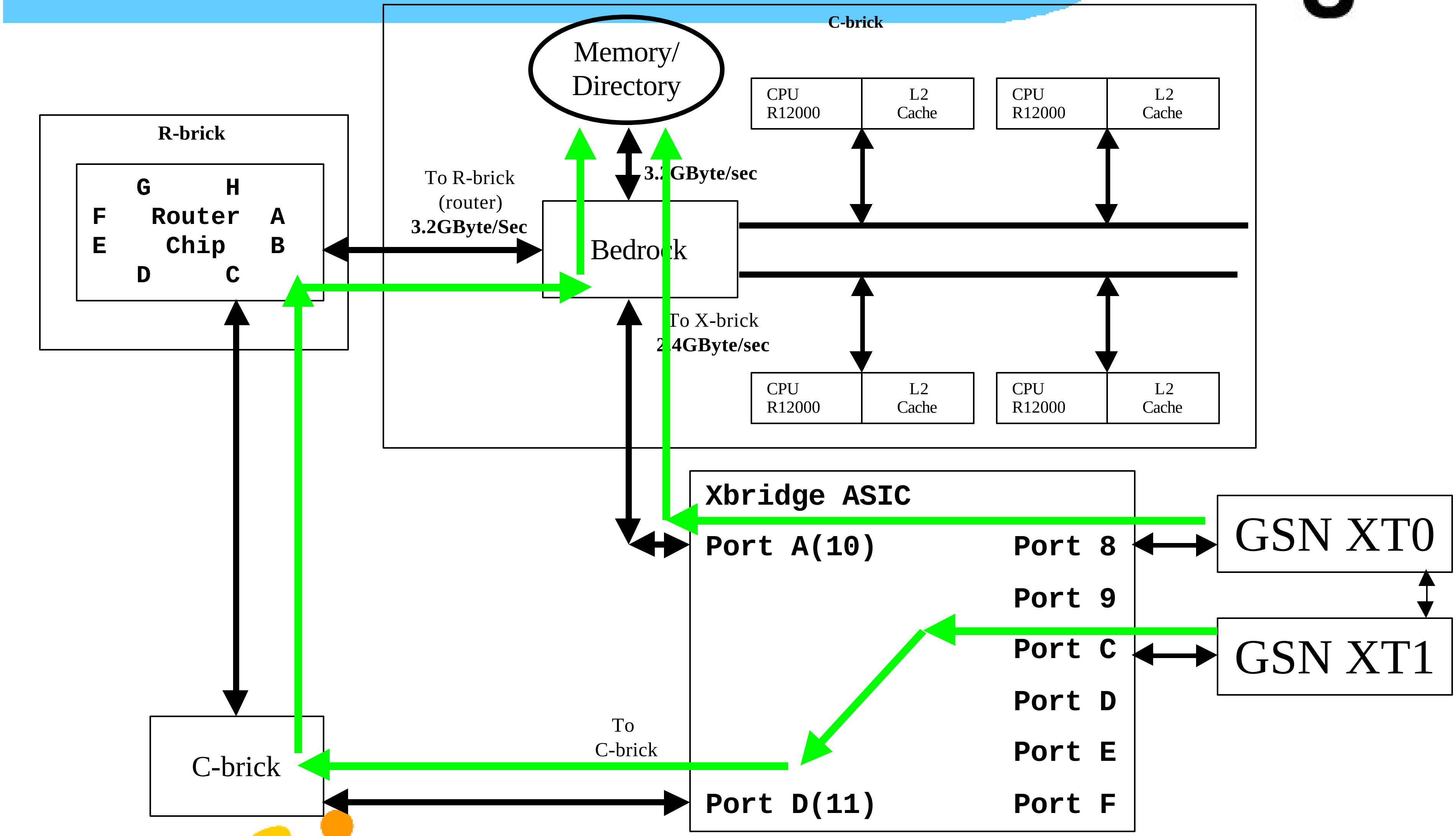
Origin 3000 Architecture

GSN Card



Origin 3000 Architecture

GSN Data Path



Comparison of O3000 vs. O2000

-*Origin 2000 performance has improved by over 25%-50% due to software improvements*

-*Origin 3000 performance numbers are showing the strength of the improved memory bandwidth of the NUMAFlex architecture*

Comparison of O3000 vs. O2000

Recent Origin 2000 performance numbers

- 250+ MB/s TCP/IP (single ttcp stream)
- 500+ MB/s ST (single memory, single stream)
- 690+ MB/s ST (2-way memory stripe, single stream)
- 791+ MB/s ST (STP diagnostic software)
- 6 µs user latency (5m cable, no switch)
- 9 µs user latency (50m cable plus switch)
- 1.45 million packets per second

Comparison of O3000 vs. O2000

Recent Origin 3000 performance numbers

-*645+ MB/s SN1 STP, 1 memory, 1 stream, 1MB pages*

-*Netperf of 610 MB/s*

Comparison of O3000 vs. O2000

Origin 3000 or Origin 2000 comparisons

Single Thread:

O3k to O3k running ST Protocol 645 MB/s

O3k to O2k running ST Protocol 613 MB/s

O2k to O2k running ST Protocol 539 MB/s

Over 100 MB/s improvement and less regard with memory placement

(gsnsttest -b96m -l96m -p20 [-s4 on O2k for O3k to O3k testing only])

Comparison of O3000 vs. O2000

Origin 3000 or Origin 2000 comparisons

Running in loopback using single path to memory and on O3k using a single cpubus (intentionally creating a worst case)

O3k to O3k on same path 360 MB/s

O2k to O2k on same path 250 MB/s

(o3k runon {4,5} gsnsttest -b96m -l96m -p20)

(o2k runon {2,3} gsnsttest -b96m -l96m -p20)

Current News

-*GSN OSBypass (libst) running MPI Apps in Beta*

-*GSN 2.0 release Sept.2000*

- **Available in single ($15K list) & dual ($25K list) XIO versions**

Review

-*SGI new NUMAFlex machine Origin 3000 running GSN*

-*Origin 3000 double the bandwidth of Origin 2000*

-*Origin 2000 great performance improvements over the past year*

-*Origin 3000 showing single adapter improvements of 100 MB/s+ over Origin 2000*

Sources for more information

Lots of info now available

- GSN Insight books
- man pages on gsn, gsn tools, stp, bds, libst, etc.
- www.hippi.org for ANSI specs
- http://oss.sgi.com/projects/stp/
- www.sgi.com/peripherals/networking/gsn_overview.html

Connectivity comparison

Technology	Bandwidth (Mbps)	Latency (us)	CPU util
GSN	6400	< 10 us	< 10%
GigE	1000	200 us	125%
Fast Ether	100	200 us	low