Building a Multicomputer with COTS Technology

Parallel multicomputers are particularly useful in solving several classes of parallelizable computational problems particularly in image and signal processing. In the past, parallel multicomputing has required custom hardware platforms with custom software development environments. Contemporary PCI-to-Switching Fabric interconnection technologies have eliminated the need for such non-standard and proprietary multicomputing platforms. Starfabric-PCI gateway technology now and Advanced Switching PCI-Express in the near future make it possible to build powerful multicomputers with standard personal computers that run standard operating systems like Linux, VxWorks or Windows.

Introduction

Parallel multicomputers are particularly useful in solving several classes of parallelizable computational problems particularly in image and signal processing. In the past, parallel multicomputing has required custom hardware platforms with custom software development environments. Contemporary PCI-to-Fabric interconnection technologies have eliminated the need for such non-standard and proprietary multicomputing platforms. Starfabric-PCI gateway technology now and Advanced Switching PCI-Express in the near future make it possible to build powerful multicomputers with standard personal computers that run standard operating systems like Linux, VxWorks or Windows.

Mercury (http://www.mc.com) is a typical supplier of custom hardware and software suitable for solving parallelizable problems including searching image databases (typically for security applications and similar tasks), computerized speech recognition, processing image data (typically in military or medical applications), nuclear simulations and meteorology. Such proprietary platforms are no longer necessary because the TTI Starfabric/Advanced Switching PCI Express Multicomputer Control Software (MCS) makes it possible to build a Multicomputer with COTS (Commercial-Off-The-Shelf) computer platforms and COTS Starfabric or Advanced Switching PCI Express adapter cards.

Developers of multicomputer application software no longer need learn the intricacies of a custom operating system API on a custom hardware platform because multicomputer applications designed for the MCS-based COTS Multicomputer are Linux, VxWorks or Windows programs that use standard Linux, VxWorks or Windows device driver system call interfaces (like *open(), read(), write(), ioctl(),* etc.).

Starfabric and Advanced Switching PCI Express

Starfabric and Advanced Switching PCI Express technologies extend the PCI bus and enable interconnection topologies beyond sequential busses or trees. The advanced switching functionality of these technologies makes it possible for interconnected host computers to access each other's host memory. Starfabric or Advanced Switching PCI Express fabrics can interconnect standard PCI host computers to create Non-Uniform Memory Access (NUMA) multicomputers.

Descriptions of such switching fabric interconnection often employ comparisons with TCP/IP computer networking, but standard LAN or WAN networked systems, which usually constitute examples of loosely coupled No Remote Access Memory Access (NORMA) architecture multicomputers, are actually very different from fabric-interconnected multicomputing systems. TCP/IP software generally can only simulate direct memory access between peer computers while Starfabric and Advanced Switching PCI Express provide this functionality in reality.

In the typical LAN technology, a bus mastering DMA device reads data from host memory (by posting an address in the first bus transaction and then receiving the data in the next

transaction), transmits packetized data to a peer device, which then writes the packet into the peer host memory. In a NUMA architecture, the writer typically arbitrates for the bus and then posts a write transaction consisting of address and data, i.e., a direct write into (remote) memory. Data transmission in a NUMA architecture system is generally far more efficient and potentially of far higher performance.

Such efficiency and higher performance do not come for free. LAN technologies typically provide simple mechanisms of addressing, path selection, data broadcast, and (sometimes non-intuitive) connection set up. In contrast, fabric technologies typically require a fairly complex configuration of several layers of hardware translation tables. The Stargen Starfabric drivers and the TTI MCS hide these details from the user.

Memory Connection Setup

Keeping track of memory channels and PCI resources requires either central allocation intelligence or a fairly complex system of hardware locking over long time periods. The Starfabric drivers allow only one host in a Starfabric network to perform memory connection setup (one can analogize this restriction to the PCI restriction that only one processor can carry out PCI bus configuration and enumeration). The MCS package provides drivers and background demons/applications that transparently and unobtrusively carry out this memory connection setup in a distributed environment.

The numa logical devices associated with the MCS numa device driver create a uniform view of local memory connections to remote devices. When an application opens up the local numa device (currently /dev/numa[0-MAXDEVICES] under the Linux operating system) for reading, the MCS creates the buffer and buffer control memory connections, and then the remote writer may start writing the local memory buffer associated with the local numa device.

The remote writer can be a true physical I/O device, or it may be an application on a remote Starfabric host that opens one of its own numa devices for writing in physical device simulator mode.

After the local reader opens up a local numa device and starts reading the data that the remote device has written into the local numa memory buffer, another reader application may be started on another Starfabric or Advanced Switching PCI-Express host. If the new reader opens up the local numa device that corresponds to the numa device that the first reader opened up on its host computer, the data that the first reader reads will be forwarded to the corresponding numa memory buffer on second reader's host machine once the MCS creates memory connections between the two machines where the readers are running. In this way a single data stream from a physical or simulator I/O device can be duplicated and distributed among many readers on many hosts within the Starfabric or Advanced Switching PCI-Express multicomputer.

Physical Input Devices

Physical input devices may either be intelligent native Starfabric (or in the future Advanced Switching PC-Express) adapters whose downloaded or built-in control software understands Starfabric (or in the future Advanced Switching PC-Express), or this type of adapter may be an intelligent PCI or cPCI adapter card whose downloaded or built-in control software understands Starfabric (or in the future Advanced Switching PC-Express) in a Starfabric-based (or in the future Advanced Switching PC-Express) expansion chassis.[1]

A device specific PCI legacy mode driver detects and downloads the Starfabric or advanced switching aware software into the intelligent adapter card. As long as the downloaded device control software understands the numa buffer and numa buffer control structures and uses them as specified in the document entitled "Writing Starfabric or PCI-Express Advanced Switching Aware Control Software for Intelligent Adapter Cards," the intelligent adapter card will be

- able to serve transparently as an input device in the COTS Multicomputer that uses the MCS and will be
- sharable or dynamically allocatable among the individual computers that comprise the COTS Multicomputer.

With custom modifications of the MCS and some extension software that can emulate fabric WMEs (Write Message Events), standard unintelligent PCI adapters may be used as input devices for the COTS Multicomputer and shared among the individual host computers within the multicomputer.

Using the System

A typical COTS Multicomputer Application would allocate the physical I/O devices to parallel pipelines of host processors within the multicomputer. The first row of host computers in the parallel pipelines would be running subunits of the parallel application that would open the physical devices, which would then write the collected data into the local memories of these host processors. The first row of host processors would then process the data and write the results into the memories of the host processors in the second row of the parallel pipelines once these computers open up the numa devices that correspond to the first row of host computers that have opened up local numa devices for writing in physical device simulator mode. According to the nature of the problem that the application software is solving intermediate results can be combined or fanned out to yet another level

[1] Carlo Gavazzi Computing Solutions sells the XP-SB Starfabric-based PCI expansion chassis product line (see http://www.gavazzi-computing.com/displayproduct.php?TGP=System%20Expansion&TGT=StarFabric). Hartmann Elektronik sells PCI and cPCI expansion chassis as well as components from which such chassis can be constructed (see http://www.hartmann-elektronik.de/en/html/en-frame.htm). Contec Corporation provides similar products (see http://www.contecusa.com/?page=prod_l2&id=58&cat_id=49&CFID=68376&CFTOKEN=28376986). Other Stargen partners (http://www.stargen.com/partners/index.shtml) produce a wide range of Starfabric-enabled technology.

of pipelines or sent back for reprocessing at earlier stages of the parallel pipelines. In this way, the TTI MCS can make possible the construction of extremely powerful multicomputer systems

- that can host the next generation of interactive video games,
- that are optimal platforms for performing efficient
    1. image processing,
    2. speech recognition, and
    3. signal processing in general, and
- that will carry out parallelizable simulations or approximation calculations in
    1. meteorology,
    2. pharmaceutical development/biotechnology,
    3. chemical syntheses,
    4. particle physics,
    5. nuclear reactions,
    6. military tactics,
    7. logistic optimization,
    8. financial modeling, etc.

Contact

For more information contact Joachim Martillo, President, Telford Tools, Inc., 617-448-6703.

Starfabric Links

http://www.stargen.com/partners/index.shtml

http://www.compactpci-systems.com/columns/starfabric_watch/pdfs/3.2004.pdf

http://www.compactpci-systems.com/columns/starfabric_watch/pdfs/11.2004.pdf

http://www.techonline.com/community/ed_resource/feature_article/32057

Advanced Switching PC-Express Links

http://www.commsdesign.com/story/OEG20021212S0019

http://www.commsdesign.com/design_corner/OEG20030410S0020

http://www.intel.com/technology/pciexpress/devnet/AdvancedSwitching.pdf

http://www.stargen.com/technolo gy/adv_switching_articles.shtml

http://www.rtcmagazine.com/home/article.php?id=100273

http://www.stargen.com/technology/adv_switching_tutorials.shtml