HyperShare™

New Industry Standard Paradigm for Data Center Computing

Hardware and Software Brief

Emilio Billi – Chair Person, HyperShare Ecosystem
Mario Cavalli – General Manager
HyperTransport Technology Consortium

Introduction

HyperTransport® with advanced node addressability (HyperShare technology) is an innovative interconnect standard that addresses the data center, the high performance computing (HPC) and the networking technology domains. This paper provides an in depth description of the most important application of HyperShare, namely its use as a high speed interconnection network (often called a system area network) for computing clusters, data center servers and System-on-Chip (SoC) micro-server applications.

For technical details, the interested reader is referred to the material than the HyperTransport Technology Consortium (HTC) makes available to its members and, most importantly, the HyperShare standard specifications and support documents.

Background

HyperShare was created by HTC as an extension of the already successful and widely adopted HyperTransport interconnect standard and capable of delivering very advanced computing node and memory addressing features, so as to enable the best optimization – i.e. minimization - of cluster computing hardware resources, thus resulting in the lowest possible system Total Cost of Ownership (TCO). Such kind of technology advancement was to specifically support the increasing deployment of scalable multi-processor, multi-node server clusters with the introduction of cross-cluster global address space and memory mapping architecture made possible by an innovative and, latency wise, very efficient software protocol governing the cluster network.

The resulting HyperShare core specification – i.e. HyperShare High Node Count Specification - released by HTC in 2009, describes the HyperShare hardware and protocol features that empower system processors in each and every clustered server (cluster nodes) with the ability to have a shared memory “view” of the entire cluster. HyperShare also specifies data read and write transactions, lock memory locations without software protocol involvement, as well as message transmission and interrupts.

In contrast to other interconnect technologies, the HyperShare interconnect is based exclusively on point-to-point links only, operating under a distributed shared memory (DSM) architecture in hardware.[1][2]

Performance

As discussed in the background section, the HyperShare standard was created with targets of minimized TCO and low latency performance for highly scalable data center, cloud computing and HPC infrastructure platforms. The interconnect network performance required to meet the goals of such distributed computing applications must be capable of:

- High sustained throughput[3]
- Low latency[4]
- Low CPU overhead for communication operations

with link bandwidth in the multi GBit/s range and latency in the low microseconds range in loosely coupled systems, with even less latency in tightly coupled multi-processor systems.

Scalability

The HyperShare industry standard was specifically created to satisfy a number of scalability requirements. Namely in:

1) Scalability of performance, as the number of clustered nodes attached to the system grows;
2) Scalability of interconnect distance, from inches to tens or even hundreds of meters - depending on the connection medium and physical layer implementation – and yet based on the same logical layer protocols;
3) Scalability of the memory system, which must not have a built-in limit on the number of processors or memory modules that could be handled (64bit of addressability)
4) Technological scalability, i.e. use of the same mechanisms in large-scale and small-scale as well as tightly-coupled and loosely-coupled systems, and the ability to readily make use of advances in technology, e.g., high-speed links;
5) Economic scalability, i.e., use of the same mechanisms and components in low-end, high-volume and high-end, low-volume systems, opening the possibility to leverage the economies of scale of mass production hardware;

[1] **Distributed Shared Memory** (DSM), in Computer Architecture is a form of memory architecture where the physically separate memories can be addressed as one logically shared address space. Here, the term **shared** does not mean that there is a single centralized memory but **shared** essentially means that the address space is shared, same physical address on two processors refers to the same location in-memory.
Patterson, David A. and John L. Hennessy (2007). *Computer architecture : a quantitative approach*, Fourth Edition, Morgan Kaufmann Publishers, p. 201. ISBN 0-12-370490-1.

[2] DSM has been implemented in many interesting hardware architectures. Some examples are: Quadrics's QsNET, Cray's Cary T3E MPP, Dolphin's PCI and 2D-3D SCI adapters, Myricom's Myrinet interconnect.

[3] High sustained throughput is intended not only in terms of pure bandwidth, but more specifically of number of packets per seconds (e.g. small packet efficiency) and numbers of concurrent data flows (e.g. virtualization flows).

[4] The objective is the minimum possible latency in a point to point communication.

6) Addressing latitude – no predictable limits to future scaling addressing requirements by means of a DSM capability wide enough to support any number of cluster nodes and a large memory capacity in each or any of them.

Topology Independence

HyperShare enables network topologies of any level of complexity. However, it is realistic to expect its initial popularity to hinge on relatively simple and efficient ring type, 2D or 3D torus topologies.
For small clusters for instance (1 to 4 nodes), the preferred topology is a small ring cluster, or "ringlet". For larger cluster topologies like multidimensional, tori are the preferred choice. For large clusters, HyperShare can also support hybrid topologies, like 2D-3D clusters of conventional switched topology sub-clusters. With this approach, HyperShare brings a high level of optimized scalability, performance and TCO to clusters that conventional switched network architectures would be either incapable, or at a performance disadvantage to support.[5]

Hardware-Based Distributed Shared Memory (DSM)

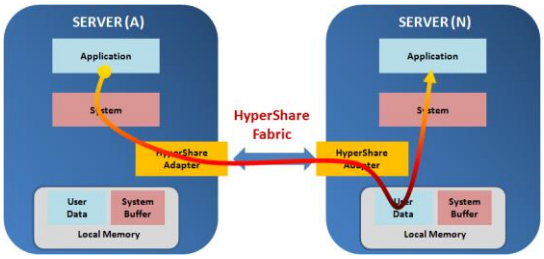
The new frontier of the "in-memory computing"

HyperShare spans a global, 64-bit memory address space; in other words, it is a physically addressed, distributed shared memory system (DSM), in which the memory distribution is transparent to the software and even to the processors, i.e. the memory is logically shared, just as in a system with a centralized bus and shared memory. This capability represents what may be considered the most important feature of HyperShare. The result is the lowest achievable communication latency.

A memory access by a processor in the cluster is mediated to the target memory module by the HyperShare hardware logic. The major advantage of this feature is that inter-node communication can be realized by simple load and store operations by the processor, without invocation of a software protocol stack.

Interconnection Network and Communication Scheme

In contrast to a LAN and most other system area networks, the HyperShare cluster network is based on memory, a physically distributed shared memory architectures, ultimately providing **a naturally fast and robust in-memory computing infrastructure without software overhead.**



I/O Subsystem Interconnect

HyperShare can be used to connect one or more I/O subsystems to a computing system in novel ways. In fact, HyperShare shared address space can include the I/O memory space and I/O nodes, which in turn are

[5] Star, Mesh, Dual-star, Hypercube, 1D-xD Torus, fat tree , and others.

enabled to directly transfer data between the peripheral devices - in most cases, storage devices - and the compute nodes' memory using DMA.

Again, software does not need to be involved in the actual transfer.

This feature speed up virtual environment like VMWARE, Cytrix, Xen, and others by providing high bandwidth and low latency direct communication between CPU and memory both local and remote one thanks to the direct remote memory access capability.

In combination with, for instance, a specific version of PCIe SRV-IO (Single Root Virtualization IO), HyperShare enables fully hardware-virtualized I/O architectures.

Server Virtualization

Server I/O Virtualization requires logical sharing of physical I/O resources between server nodes and the rest of cluster resources. In line with these objectives, HyperShare empowers the cluster infrastructure with:

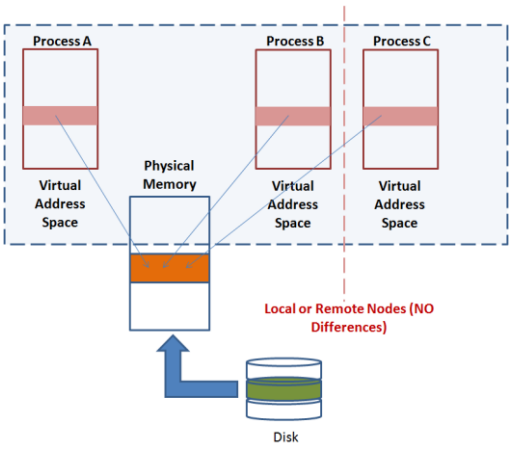
A. On-demand efficient allocation of I/O
B. Matching of CPU and OS processing requirements to I/O
C. Better CPU utilization
D. I/O resource sharing between all cluster nodes
E. Memory virtualization

Memory-Mapped Files

File memory-mapping allows a process to access the content of a file directly in-memory through normal load and store instructions. This is the perfect in-memory computing platform for databases and IO intensive applications.

Memory-mapped files can be used to directly access remote physical memory in a HyperShare cluster.

Direct access to remote memory may not have the same semantics as the access to local physical memory.



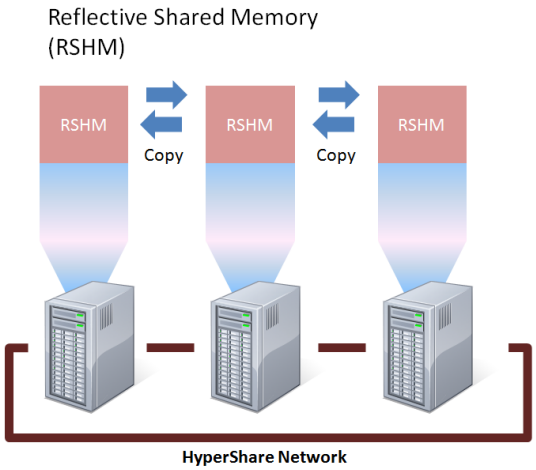
Example of HyperShare DSM architecture operation

Advanced use of HyperShare shared memory architecture and its related hardware supports for DSM

We would like to introduce a special application of HyperShare DSM, called Reflective Memory[6] (RM).

RM systems (in computer literature also referred to as mirror memory systems, replicated shared memory, multicast or replicated memory systems) implement transparent and automatic updates of remote shared memory areas. Network architectures based on this specific memory configuration provide the tightly timed performance necessary for all kinds of distributed simulation, industrial control applications and military applications.

Since their inception, RM networks have benefited from advances in general purpose data networks, but they remain an entirely independent technology realm, driven by different needs and catering to a different set of users.



RM can be best leveraged in :

1) Clustering
2) Supercomputing
3) Distributed Computer Systems

As stated above, an RM network is a special type of shared memory system, designed to enable multiple, individual computers to share a common set of data. Unlike global shared memory systems, in which individual systems is able to access a single, memory space, RM networks place an independent copy of the entire shared memory set in each attached system.

In combination with HyperShare hardware DSM, RM enables real time point to point and point to multipoint communications in a very unique way. Each attached system has full, unrestricted rights to access and change this set of local data at the full speed of writing to local memory.

Today RM is used in hundreds of applications to network computers together. Specifically, any application for which the designer desires (or requires) high performance or ease of use (or both), RM is the ideal choice.

RM is currently utilized in the following applications:

- Aircraft simulators
- Automated testing systems
- Ship and submarine simulators
- Power plant simulators
- Industrial process control
- High speed data acquisition

[6] A Reflective Memory network is a Real Time Local Area Network that offers unique benefits to the network designer. Reflective Memory has become a de facto standard in demanding applications where determinism, implementation simplicity, and lack of software overhead are key factors.

HyperShare can support RM in hardware thanks to it 64 bit memory addressing and its DSM architecture enabling powerful software application based on it.

Implementing RM software can be considerably easier than creating a complete, traditional DSM architecture. In this scenario, HyperShare can play a true leadership role in enabling that creation of new-generation cluster systems hundreds of times faster than traditional software-based ones.

RM-based computing systems are particularly well suited for building large scale, fault-tolerant disk storage and file systems for I/0 intensive applications.[7]

HyperShare and Ethernet

Using the HyperShare–over-Ethernet (HSoE) capability supported by the HyperShare protocol specification, all above capabilities can as well be implemented and leveraged in a standard Ethernet environment. This means that HSoE operates over standard Ethernet switches and cables, thereby empowering conventional Ethernet hardware infrastructures with DSM capability.

The HyperTransport Technology Consortium has standardized and released a dedicated HyperShare-over-Ethernet specification supporting 10, 40 or 100 Gigabit HSoE implementations.

Conclusions

The most noticeable and distinctive feature of HyperShare is its shared memory capability, i.e., the transparent memory access by remote compute nodes. By hinging on its DSM strength, HyperShare can provide orders of magnitude better performance than any other network interconnect technology. By also supporting MPI and TCP/IP protocols over its shared memory architecture, HyperShare delivers the much sought after benefits of in-memory computing architectures to every kind of data center application, with major reduction in power consumption and TCO and increased overall data center efficiency.

[7] "Implementation of a Fault-Tolerant Disk Storage System Using Reflective Memory" : Nicos Vekiarides, Department of Electrical and Computer Engineering, Carnegie Mellon University