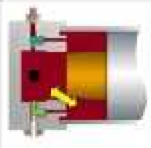
electronic design The Authority on Emerging Technologies for Design Solutions

Enter Search Term GO DRILLDEEPER: Enter Drill Deeper or ED Online ID GO

over 100 Design VIDEOS

Technologies Design Hotspots Resources Shows Magazine eBooks & Whitepapers Jobs Mor

Click to view this week's ad screen



[Design View / Design Solution]

Break Through The TCP/IP Bottleneck With iWARP

Using a high-bandwidth, low-latency network solution in the same network infrastructure provides good insurar generation networks.

Sweta Bhatt, Prashant Patel

ED Online ID #21970
October 22, 2009

The online economy, particularly e-business, entertainment, and collaboration, continues to dramatically and rapidly increase the amount of Internet traffic to and from enterprise servers. Most of this data is going through the transmission control protocol/Internet protocol (TCP/IP) stack and Ethernet controllers.

As a result, Ethernet controllers are experiencing heavy network traffic, which requires more system resources to process network packets. The CPU load increases linearly as a function of network packets processed, diminishing the CPU's availability for other applications.

Because the TCP/IP consumes a significant amount of the host CPU processing cycles, a heavy TCP/IP load may leave few system resources available for other applications. Techniques for reducing the demand on the CPU and lowering the system bottleneck, though, are available.

iWARP SOLUTIONS

Although researchers have proposed many mechanisms and theories for parallel systems, only a few have resulted in working cor One of the latest to enter the scene is the Internet Wide Area RDMA Protocol, or iWARP, a joint project by Carnegie Mellon Univer Corp.

This experimental parallel system integrates a very long instruction word (VLIW) processor and a sophisticated finegrained commu on a single chip. It's basically a suite of wireless protocols comprising RDMAP and DDP. The iWARP protocol suite may be layere PDU aligned (MPA) and TCP or over Stream Control Transmission Protocol (SCTP) or other transport protocols (*Fig. 1*).

The RDMA Consortium released the iWARP extensions to TCP/IP in October 2002, implementing the standard for zerotransmissic TCP/IP. Together, these extensions eliminate the three major sources of networking—transport (TCP/IP) processing, intermediate application context switches—that collectively account for nearly 100% of CPU utilization (*see the table*).

A kernel implementation of the TCP stack has several bottlenecks. Therefore, a few vendors now implement TCP in hardware. Be loses are rare in tightly coupled network environments, software may perform the error-correction mechanisms of TCP. Meanwhile on the network interface card (NIC) strictly handles the more frequently performed communications. This additional hardware is kn offload engine (TOE).

The iWARP extensions utilize advanced techniques to reduce CPU overhead, memory bandwidth utilization, and latency. This is a through a combination of offloading TCP/IP processing from the CPU, eliminating unnecessary buffering, and dramatically reducin operating-system (OS) calls and context switches. Thus, the data management and network protocol processing is offloaded to an Ethernet adapter instead of the kernel's TCP/IP stack.

iWARP COMPONENTS

Offloading TCP/IP (transport) processing: In conventional Ethernet, the TCP/IP stack is a software implementation, putting a treme host server's CPU. Transport processing includes tasks such as updating TCP context, implementing required TCP timers, segme reassembling the payload, buffer management, resource-intensive buffer copies, and interrupt processing.

The CPU load increases linearly as a function of the network packets processed. With the tenfold increase in performance from 1- 10-Gigabit Ethernet, packet processing and the CPU overhead related to transport processing increases up to tenfold as well. In th processing will cripple the CPU well before reaching the Ethernet's maximum throughput.

The iWARP extensions enable the Ethernet to offload transport processing from the CPU to specialized hardware, eliminating 40% overhead attributed to networking (Fig. 2). The transport offload can be implemented by a standalone TOE or be embedded in an Ethernet adapter that supports other iWARP accelerations.

Moving transport processing to an adapter also rids a second source of overhead—intermediate TCP/IP protocol stack buffer copi these copies from system memory to the adapter memory saves system memory bandwidth and lowers latency.

RDMA techniques eliminate buffer copy: Repurposed for Internet protocols by the RDMA Consortium, Remote DMA (RDMA) and I Placement (DDP) techniques were formalized as part of the iWARP extensions. RDMA embeds information into each packet that application memory buffer with which the packet is associated. This enables the payload to be placed directly in the destination ap even when packets arrive out of order.

Data can now move from one server to another without the unnecessary buffer copies traditionally required to "gather" a complete This is sometimes called the "zero copy" model. Together, RDMA enables an accelerated Ethernet adapter to support direct-memo from/writes-to application memory, eliminating buffer copies to intermediate layers. RDMA and DDP eliminate 20% of CPU overhe networking and free the memory bandwidth attributed to intermediate application buffer copies.

Avoiding application context switching/OS bypass: The third and somewhat less familiar source of overhead, context switching, co significantly to overhead and latency in applications. Traditionally, when an application issues commands to the I/O adapter, the co transmitted through most layers of the application/OS stack.

Continue to page 2

Passing a command from the application to the OS requires a CPU intensive context switch. When executing a context switch, the the application context in system memory, including all of the CPU general-purpose registers, floating-point registers, stack pointer pointer, and all of the memory-management-unit state associated with the application's memory access rights. Then the OS contex loading a similar set of items for the OS from system memory.

The iWARP extensions implement OS bypass (user-level direct access), enabling an application executing in user space to post c to the network adapter (Fig. 4). This eliminates expensive calls to the OS, dramatically reducing application context switches. An a Ethernet adapter handles tasks typically performed by the OS. Such adapters are more complex than traditional non-accelerated N eliminate the final 40% of CPU overhead related to networking.

POTENTIAL APPLICATIONS

Clustered servers/blades depend on high-bandwidth, low-latency interconnects to aggregate the enormous processing power of do and keep I/O needs serviced. For these applications, iWARP is advantageous because it delivers:

- Predictable, sustained, scalable performance, even in multicore, multi-processor clusters
- A single, self-provisioning adapter that supports all clustering MPIs: HP MPI, Intel MPI, Scali MPI, MVAPICH2
- Modern low-latency interconnect technologies to an ultra-lowpower, high-bandwidth Ethernet package; flexibility, cost, and indust management benefits; and exceptional processing power per square foot for clustered applications.

For data-networking applications, iWARP-based accelerated Ethernet adapters offer full-function data-center NIC capabilities that performance, improve power efficiency, and more fully utilize data-center assets. The accelerated Ethernet adapters achieve the h lowest latency, and lowest power consumption in the industry.

In particular, when it comes to network overhead processing, it achieves 95% CPU availability for the application and operating sys hardware-class performance for virtualized applications ensures offloaded CPU cycles from network processing are not lost to soft overhead.

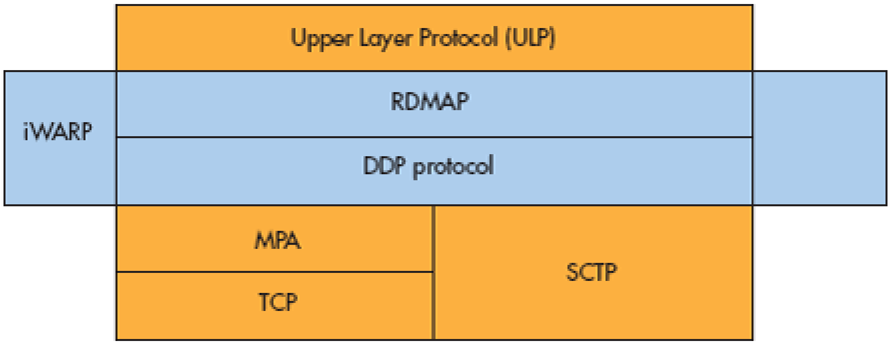
For storage vendors, iWARP will become a standard feature of network-attached storage (NAS) and Internet small computer syste (iSCSI) storage access networks (SANs). NAS and iSER-based (iSCSI Extensions for RDMA) storage networks utilizing accelerat adapters deliver the highestperformance, lowest-overhead storage solutions at any given wire rate.

At 10 Gbits/s, Ethernet-based storage networks offer a viable alternative to a Fibre Channel SAN. There's a single network interfac file protocols, high-throughput block, and file-level storage access, exceeding 8-Gbit Fibre Channel.

In addition to iSCSI, iWARP supports a wide range of other storage protocols such as network file system (NFS) and common Inte (CIFS). This gives data centers the opportunity to reap the increased productivity and lowered total cost of ownership benefits of a standards-based technology.

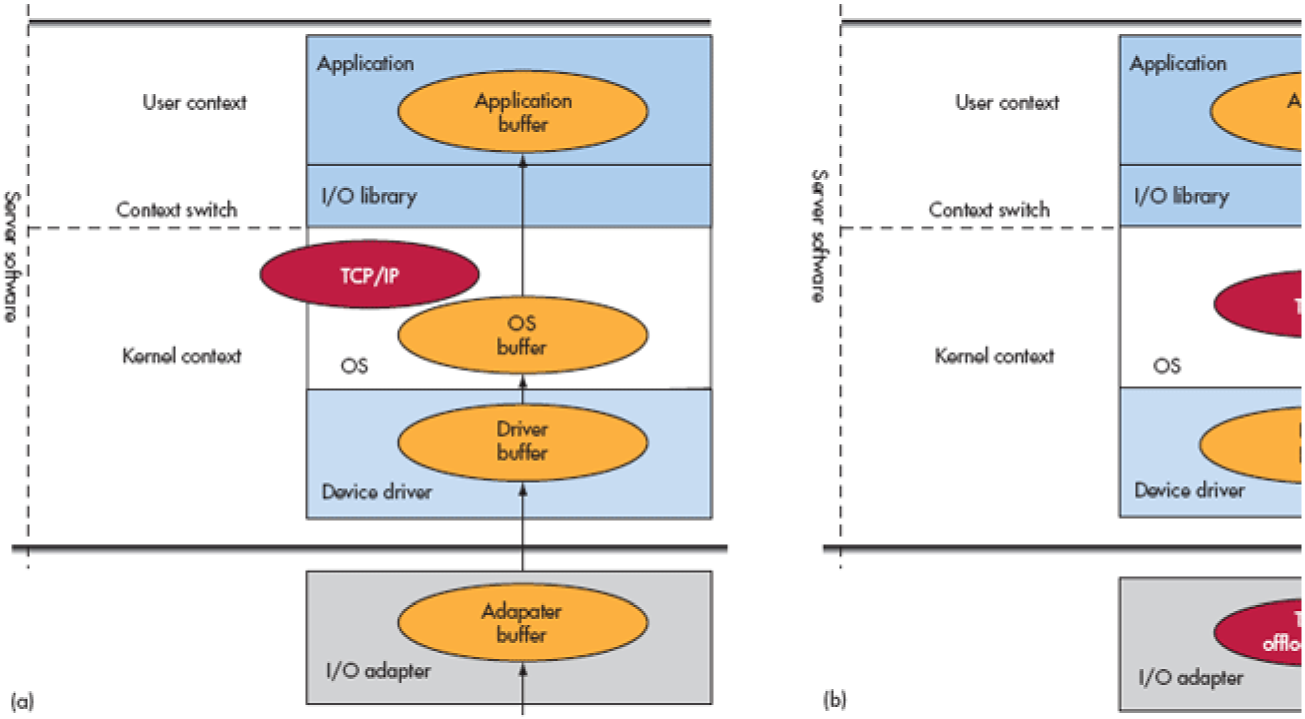
Like InfiniBand, iWARP does not have a standard programming interface, only a set of verbs. Unlike the InfiniBand architecture (IE has reliable connected communication as this is the only service that TCP and SCTP provide. The iWARP specification also omits special features of IBA, such as atomic remote operations. In all, iWARP offers the basics of InfiniBand applied to Ethernet. This su legacy software and next-generation applications.

fig 1



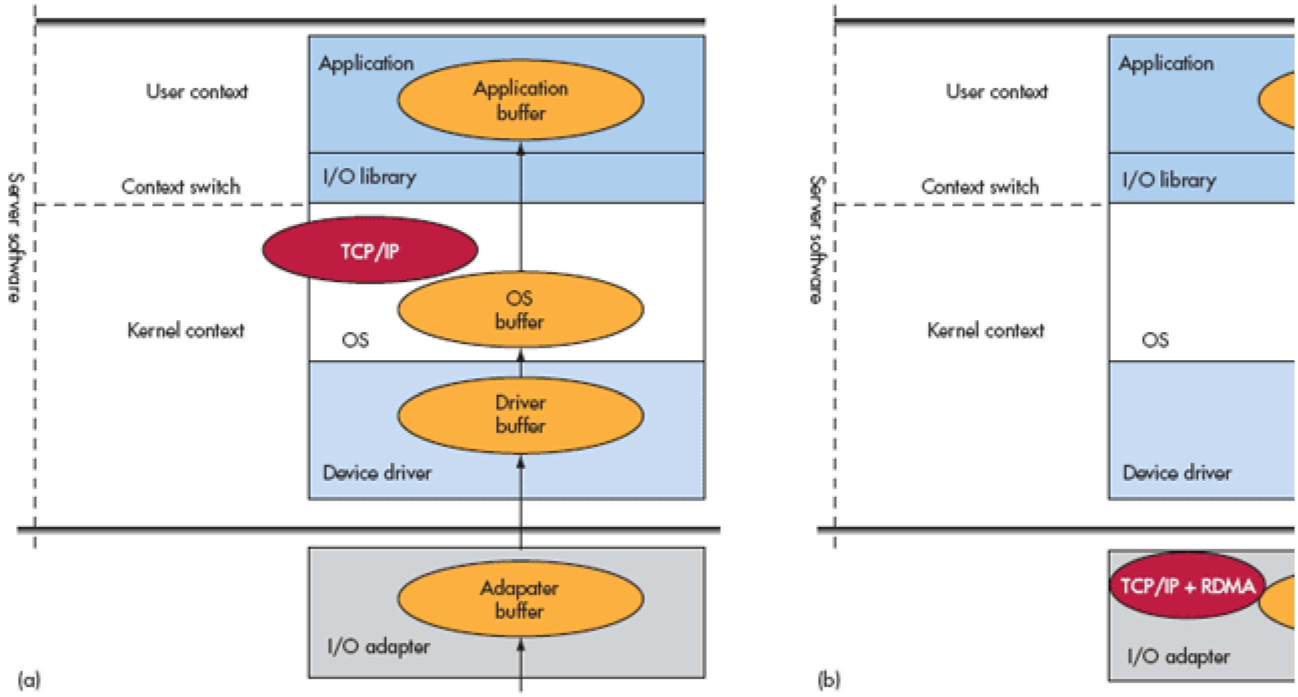
1. iWARP can be layered above MPA over TCP or over SCTP or other transport protocols.

fig 2



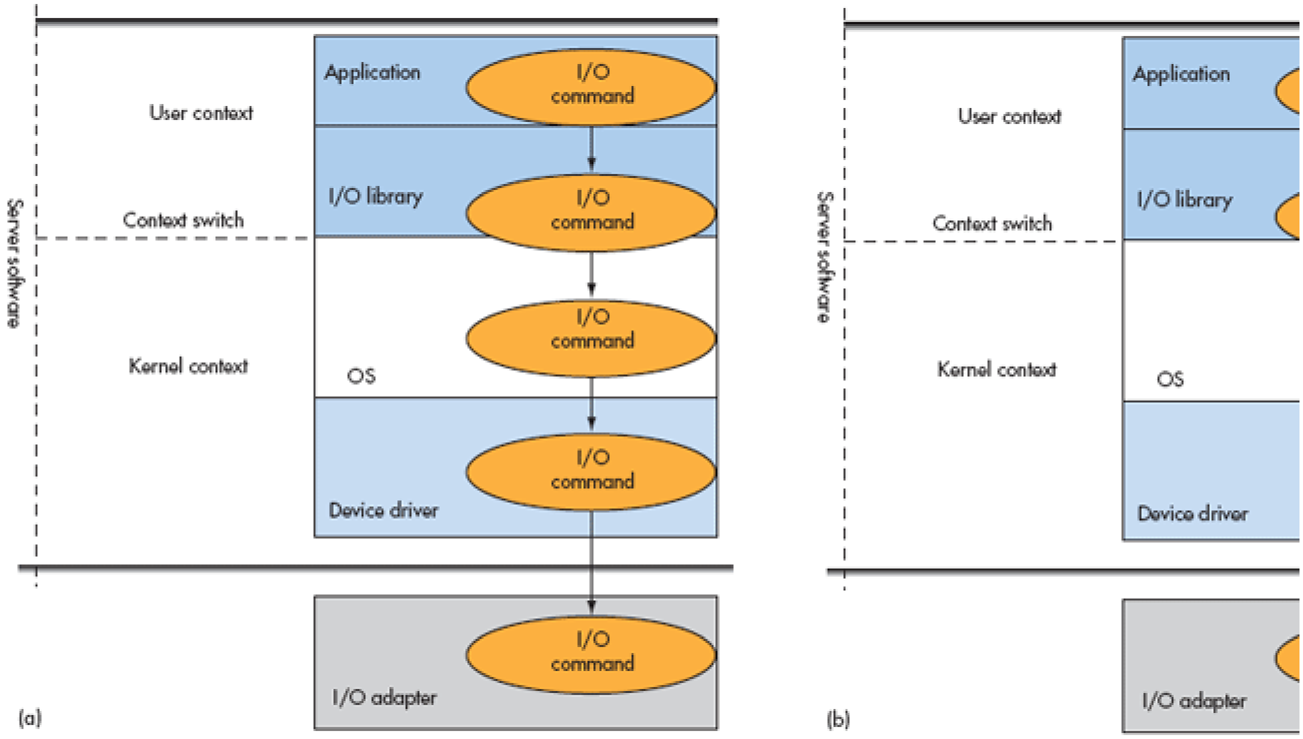
2. As opposed to a standard TCP stack (a), one that uses a TCP/IP offload engine can reduce CPU overhead by 40% (b).

fig 3



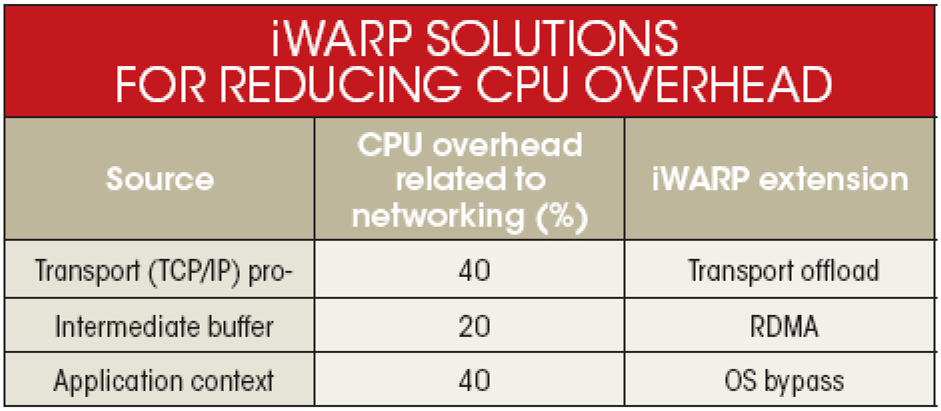
3. Shown is a standard TCP/IP stack without RDMA (a) and a stack containing Ethernet packets with RDMA that eliminate buffer layers, reducing overhead (b).

fig 4



4. The diagram on the left illustrates context switching with standard OS calls (a). An iWARP-enabled adapter can institute OS the adapter handles OS tasks and eliminates otherwise expensive calls to the OS.

table 1



iWARP SOLUTIONS FOR REDUCING CPU OVERHEAD

Source	CPU overhead related to networking (%)	iWARP extension
Transport (TCP/IP) pro-	40	Transport offload
Intermediate buffer	20	RDMA
Application context	40	OS bypass

Electronic Design Europe Electronic Design China EEPN Power Electronics Auto Electronics Microv
Mobile Dev & Design Schematics Find Power Products Military Electronics EE Events Related Re

Planet EE Network | Contact Us | Editorial Calendar | Media Kit | Submit Articles | Headlines | Subscribe
Advanced Search | Help | Site Map | E-mail Webmaster