The Quest for the Perfect Server for Network Computing Applications

Pankaj Mehra
NonStop Labs
HP
Cupertino, CA, USA
pankaj.mehra@hp.com

Abstract — This talk begins with an examination of server-side architecture and workload trends. Severe performance and availability gaps in server architecture, compounded by the growing cost and complexity of development and deployment, are shown hampering the successful adoption of web services. Two architectural remedies are proposed: (I) unification of I/O and memory; and (II) whole system programming. The proposed Apsara architecture for servers combines these new techniques with the best of existing server architecture concepts. The talk concludes with identification of key technical challenges and opportunities, and insights from early experiences in implementing Apsara.

Keywords — memory hierarchy, web services, computer architecture, input/output (I/O).

I. WEB SERVICES

Web services is the architectural concept of choice for developing business applications for heterogeneous, distributed computing. Its architectural foundations consist of:

- The Java language and the J2EE portability platform APIs: Java’s managed memory model, and its pointer-rich data structures, challenge the memory subsystem; however, memory latency in systems is improving at only about 7% per year. Researchers are looking into smart memory controllers and processor-in-memory approaches to offload memory-intensive work closer to memory.

- XML and related data formats (SAML, ebXML, UBL) carried inside SOAP protocol messages transported over HTTP connections: Validating XML in Java costs almost 100 instructions/byte; yet hardware acceleration in servers has just barely started to deliver network interface devices that support TCP offload. (In contrast, unaccelerated TCP costs only about 2.5 instructions/byte.) Even though servers are slow at processing Java, Data Power and Sarvega have already announced XML routers that accelerate XML handling 100-fold.

- Distributed transactions against online data stores: The goal is to provide 1000s of transactions per second, at six 9s of availability, highest levels of data integrity, and commodity price/performance of well below the $10-15/TPM we see today for clustered systems. Memory-semantic networks, such as Remote-DMA (RDMA)-enabled Ethernet technology called iWARP, promise to lower the latency and CPU utilization of inter-processor communication between database instances. They also support low-cost file I/O through user-mode storage protocols, such as Direct Access File System (DAFS) and iSCSI Extensions for RDMA (iSER).

II. THE APSARA ARCHITECTURE

The proposed Apsara architecture makes several departures from traditional methods of server design:

1. It extends the concept of System Area Networking by making network components -- interfaces, switches and routers -- more "system like", as well as by managing and scaling the system in a more "network like" fashion.

2. It extends the concept of a programmable computer by opening up more of the system to application programs. Numerous examples of this will be shown.

3. It eliminates the distinction between memory and I/O. Three key techniques used to accomplish this are (3a.) the replacement of the conventional driver-adapter model of I/O with an "I/O as a service" model best exemplified by iSCSI; (3b) the adoption of RDMA-enabled networking; and (3c) strong use of "memory-semantic" communications across the board. Several examples of this will be shown as well.