White Paper

Enterprise Distributed Systems and Infiniband

Enterprise systems are increasingly strained by the sheer volume of data that is consumed, generated, and manipulated during the course of everyday operations. This paper discusses the challenges faced by different industries and how networking technology can help businesses scale system performance.

In almost every business sector, enterprise systems are required to manage ever-increasing volumes of data.

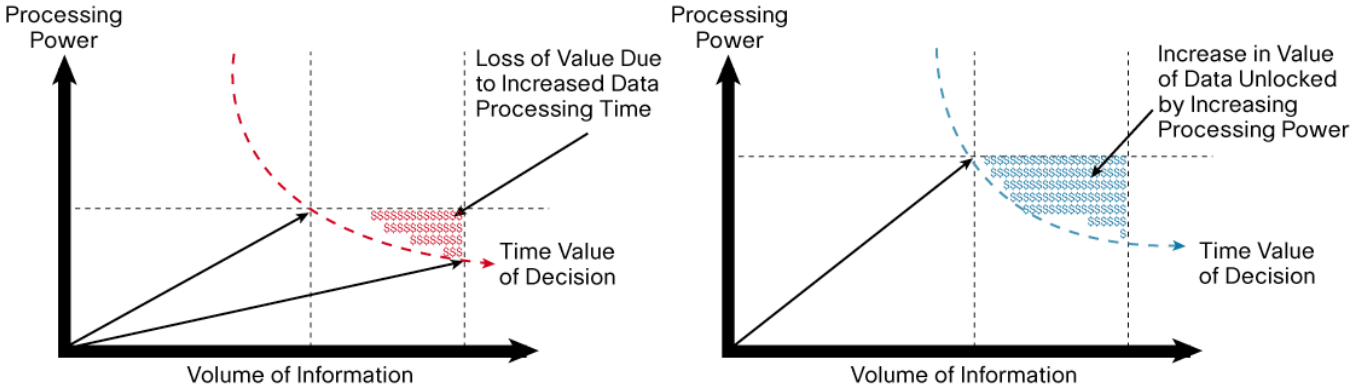
- In financial markets, the move to electronic trading fundamentally shifted stock trading to a real-time environment. Because of the volume and speed of information generated by the stock exchanges, autonomic trading systems now analyze technical indicators to buy and sell equities, where fractions of a second in delay can significantly affect the strike price of the equity.
- In the oil and gas industry, because of the rising cost of energy, companies are revisiting previously uneconomic fields to extract what resources are available. To do so, they are increasing the resolution of their seismic field analysis to get a more comprehensive view of underground energy reservoirs. Additionally, new techniques are enabling these companies to better assess new fields, or better manage existing oil or gas fields to increase production.
- The costs associated with researching and developing new drugs within the pharmaceutical industry are rising, partly due to the high rate of failure in bringing new compounds to market. Systems biology is allowing these companies to model new compounds, or to focus development on compounds that are most likely to succeed, instead of relying solely upon laboratory experimentation and clinical testing.
- The cost of developing new cars and airplanes is escalating to such a degree that the research and development costs need to be shared across multiple manufacturers. The cost-prohibitive nature of traditional techniques is compelling manufacturers to use computer systems for aerodynamic and crash modeling, in addition to computer-added design (CAD) and computer-aided manufacturing (CAM), instead of relying on physical modeling and experimentation.
- Retail organizations gather information on millions of customers to assess their demographics and buying habits, and use this information to target specific promotions to the needs of the individual customer. As systems—especially online systems—gather and store more data about an individual customer's habits, more data must be analyzed to identify trends and to increase customer retention and spending.
- Decision support systems must be able to accommodate, search, and manipulate sophisticated data to support the business and provide auditable trails for compliance and regulatory purposes.

A recurrent theme across all of these sectors is the sheer volume of data that needs to be analyzed, assessed, and used strategically, in real time, or as near real time, as possible. The mapping of the human genome provides a powerful example of how analyzing data faster can affect the competitive nature of a business. Two institutions—one private, one public—were competing to publish the exact sequence. The stakes in this exercise were very high: if the private company were able to complete the mapping first they could patent and copyright the genome sequence, and make royalties estimated to be worth between US $600 and 700 million *per year*. For the public institution, the status accrued by publishing the sequence first, and making the genome mapping publicly available and royalty-free for future drug development, were primary motivations. In the end, the public institution was able to reconfigure its high-performance compute clusters and narrowly beat the competition in mapping the human genome.

A primary element of unlocking the information that is latent within the data is the use of *distributed systems* that can process data in a timely and efficient manner. However, as the volumes of data grow and if processing capacity stays the same, the time—or latency—incurred in deriving the right information can significantly impact the ability of a company to react to market changes, especially for markets that are real time.

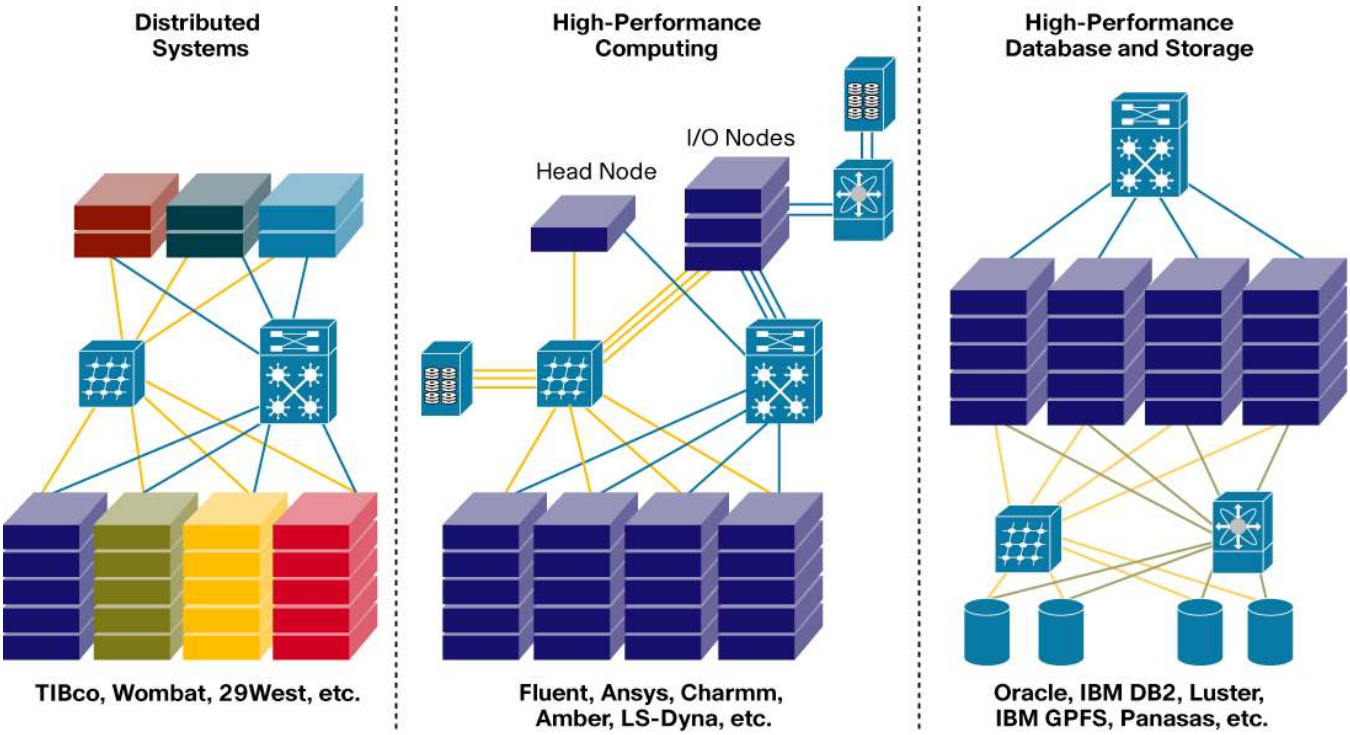
The ability to swiftly react to specific business events is critical to business. This is because the number of choices that are available and the value of a particular decision to the business *decrease* over time. The longer it takes the system to react, the more it decreases the number and value of options when addressing a particular event. Figure 1 shows how, as the volume of information increases, because the amount of processing power is static, it takes longer for the system to process the information and consequently reduces the number of choices and limits flexibility. If the processing power of the system can be increased by increasing the system performance—either by adding more or faster processors or by increasing system efficiency—the information can be processed faster, which provides more choices and greater business flexibility, agility, and efficiency.

Figure 1. The Value of Time and Data Processing



Distributed systems can be categorized as either multiple computer systems collaborating to deliver a single application, or multiple applications collaborating together as a system. The former can be broadly applied to high-performance computing (HPC) whereby multiple computers, commonly called *clusters*, collaborate to solve a single problem, or to distributed database systems that exchange information to maintain cache consistency. The latter can be broadly applied to application integration whereby multiple discrete applications receive a copy of "some" data and process the data accordingly. The primary common characteristic of these two systems is that the systems are distributed; that is, they run on multiple computers that are interconnected to form a system (Figure 2).

Figure 2. Distributed Systems Architectures

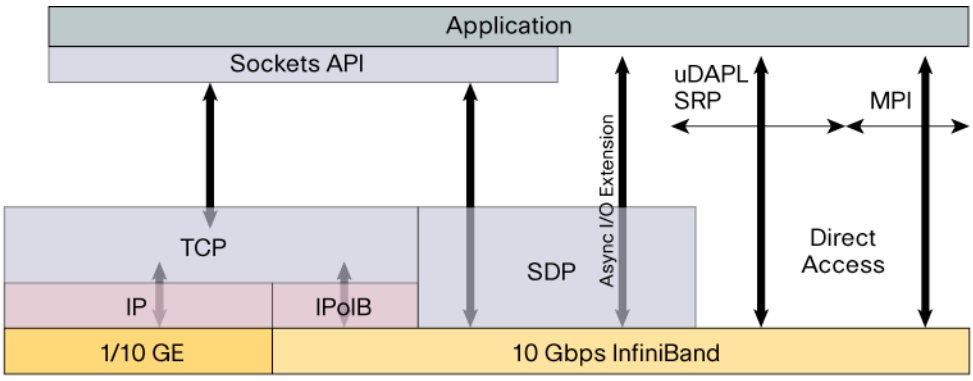


Although TCP/IP and Ethernet are widely deployed to perform this function, application developers are looking to techniques pioneered within HPC—such as InfiniBand, remote direct memory access (RDMA), sockets direct protocol (SDP), and SCSI RDMA Protocol (SRP)—to boost the performance of distributed systems for *specific applications*. For some classes of applications, systems latency—the time taken for a message to be transmitted from one application to another application (or applications)—is a critical metric. For these applications, InfiniBand provides a number of enhancements with respect to application and systems performance that can be significant when the entire system is taken into account:

- InfiniBand 20-Gbps bandwidth and low-latency reduces server-to-server message latency.
- InfiniBand transport functions are all implemented in hardware that offloads all communications tasks such as reliable, in-order delivery and multiplexing, which allows more CPU cycles to be spent on processing rather than communications.
- RDMA enables the application to offload all communications management to the InfiniBand host channel adapter (HCA), which allows more CPU cycles to be spent on processing rather than communications.

Although some benefit can be accrued by running IP natively over InfiniBand, to realize the greatest performance gains from the InfiniBand network fabric, applications need to be able to use the transport and RDMA capabilities of the InfiniBand HCAs. To use these capabilities *without* rewriting application code, the sockets direct protocol (SDP) offers a mechanism whereby SDP intercepts the application socket call within the kernel and "routes" the connection natively across the InfiniBand fabric. For those situations where latency and throughput are absolute necessities, writing the application to use SDP asynchronous I/O mode, or to use the InfiniBand interface directly, can yield further performance gains. These techniques also significantly improve the computer system's efficiency because the CPU is not involved with processing the TCP stack nor data movement. Figure 3 compares Gigabit Ethernet, Ten Gigabit Ethernet and InfiniBand transport options

Figure 3. InfiniBand Transport Comparison



Throughput	1 Gbps	3.9 Gbps	4.1 Gbps	4.5 Gbps	7.9 Gbps	8 Gbps	8 Gbs
Latency	40 usec	40 usec	20 usec	15 usec	15 usec	8 usec	4.5 usec
CPU Utilization	30%	80%	85%	25%	4%	1%	>1%

MPI: Message Passing Interface
SRP: SCSI Remote Protocol
uDAPL: User-level Direct Access Programming Language

Another popular solution is to use multiprocessor computers running a number of applications as processes within the same computer as a mini "system in a box." In effect, the computer's memory becomes the network as each individual process swaps data between different memory locations. However, this trend does not eliminate the requirement for high-speed networking because highly available processing may require that two or more systems receive the same information for processing, or exchange information to maintain the operational integrity of the application. For applications, or business imperatives, that are sensitive to latency, InfiniBand provides a high-speed, low-latency network interconnect. For those applications that are less sensitive to latency, Gigabit Ethernet or 10 Gigabit Ethernet are good choices.

SUMMARY

Enterprise systems are increasingly strained by the volume of data, and the need to rapidly assess and react to this information is becoming critically important in today's real-time enterprise. Although server technology can process vast amounts of data, business decisions are becoming more and more complex, and the use of multiple systems collaborating to solve business problems is widely adopted. Additionally, the performance of application integration systems is becoming constrained by the volume of traffic and communications overhead.

Networking technologies such as InfiniBand can be used to increase the performance of existing distributed systems, and to deliver new high-performance applications that can significantly increase competitive advantage and productivity, and reduce operational costs. InfiniBand provides high-throughput and low-latency transport for efficient data transfer between server memory and I/O devices without CPU intervention. By using protocols such as SDP, older applications can take advantage of the capabilities of InfiniBand hardware to improve system performance and efficiency.

Cisco Systems®, the world leader in networking, provides industry-leading InfiniBand technologies that power many of the world's supercomputers. The Cisco® InfiniBand portfolio includes the Cisco SFS 7000 Series InfiniBand Server Switches, Cisco SFS 3000 Series Multifabric Server Switches, Cisco SFS High-Performance InfiniBand Subnet Manager, and Cisco InfiniBand Host Channel Adapters.

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the Cisco.com Website at www.cisco.com/go/offices.

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus • Czech Republic Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel • Italy Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Printed in USA C11-360533-00 08/06