



White Paper

Commex Vulcan: In-server Content Aware Routing Network Interface Card

By Yehiel Engel

yehiele@commextech.com

Chief Architect
Commex Technologies

July 2009

Table of Contents

1 Abstract	3
2 Content Aware Routing Concept.....	3
2.1 Vulcan Architecture – Benefits 5	
3 Implementation	5
3.1 Differential Services 5	
3.2 Content Aware Bypass 6	
3.3 Network monitoring server Cluster 7	
4 Summary.....	8

1 Abstract

Conventionally, all traffic arriving from the network to a server is sent directly to the server or hardware appliance processors. The processors are responsible for classifying and directing data packets to their final destination – clients, other servers, other network agents, or storage devices. In most scenarios, the server processors only actually need to inspect and handle a small amount of data, primarily packet headers or descriptors containing packet information. Sending all the packet data to the processors causes several problems, including increased processor utilization, increased cache pollution due to unnecessary data, increased power consumption, and other computational task latency.

The now ubiquitous multicore server environment leads to more challenges. Without content aware classification at the network interface, packets may not be sent to the processing core that needs to process them. The packet may then hop from core to core until it lands at the right core. The result is unnecessary processing and cache pollution both leading to reduced performance and efficiency.

Commex patent-pending Content Aware Routing technology addresses the above challenges and is implemented in Commex Vulcan Network Interface Cards (NICs). On Commex Vulcan NICs, the implemented architecture is based on content aware classification units that perform classification on traffic on an incoming network port. These units, coupled with any-port-to-any-port routing capability and an advanced host interface based on multiple DMA channels and multiple interrupt mechanisms, allow sending only the required portion of the data to the correct target processing core. Furthermore, the architecture allows sending data to other network ports, including the source network. I/O utilization improves significantly with lower latency per packet, reduced CPU utilization and cache pollution. The result is much more efficient leveraging of the multiple processing cores on the server and higher overall server performance. Specifically, Commex Content Aware Routing technology results in multicore scalable server performance.

2 Content Aware Routing Concept

For networks, specifically those with routers and switches, content aware routing has been standard practice for many years. This is not the case for servers or hardware appliances. Commex architecture migrates several of the concepts used in network switching to the server platform. In addition, Commex integrates these concepts with an extensive and powerful host interface.

The basic concepts acquired from the networking world and implemented in Commex include:

- **Classification of incoming traffic.** Packet classification where each packet has a header and a payload (optional), and where classification is performed on the entire packet. The classification engine can be used to find specific fields (e.g., source IP, MAC ID), to locate packet formats (e.g., TCP, HTTP), or to locate specific patterns in the packet (e.g. the text string “john smith”). Based on the fields and patterns extracted, the Commex Vulcan NIC sends the relevant data to a specific host core or other network port. The classifier can also decide to send only a portion of the data

to one location and send the remaining data to another location in the host memory. This task is referred to as a "header split" - where the header is sent to further processing, while the data payload is sent to a storage buffer.

- **Any-port-to-any-port routing ability.** Commex Vulcan contains an internal switching unit that allows non-blocking of any-port-to-any port-wide communication. This mechanism utilizes high throughput and low latency features. Configurable priority queues at the edges of the interconnect unit result in improved QoS between the various types of traffic, resulting in better bandwidth management.

Vulcan host interface includes mechanisms to interact with the host multiple processing cores:

- **Multiple configurable DMA channels per RX and TX paths.** Each DMA channel can be used for different processing cores with different policy mechanisms (e.g., drop/no drop), different interrupt coalescing values per interrupt, and different interrupt vectors. A configurable scheduler is used to control the priority between these DMA engines, enabling different services to be handled according to their requirements (e.g., latency, BW).
- **Multiple interrupt vectors.** Using an MSI-X mechanism, Vulcan generates multiple interrupt messages (interrupt vector per DMA engine). Generating interrupt messages per processing allows for efficient interrupt usage, according to the application's needs. The New API or "NAPI" from Linux is a modification to the device driver packet processing framework designed to improve the performance of high-speed networking. Vulcan NAPI support results in improved system stability when there's network traffic congestion.

Figure 1 depicts Vulcan's architecture. Classification units (CLA) are used to classify incoming traffic from the network. Action Engines (AE) are then used to perform the classifier decisions (e.g., forward packet, drop, modify packet, check fields, etc). Packets are then encapsulated with an internal data structure that includes the routing information and action to be performed at the next stage. Outgoing traffic from the switch fabric is classified at a later stage, using mainly encapsulated data structure headers, while the action engine performs the relevant commands - e.g., add fields, Cyclic Redundancy Check (CRC), or time stamp.

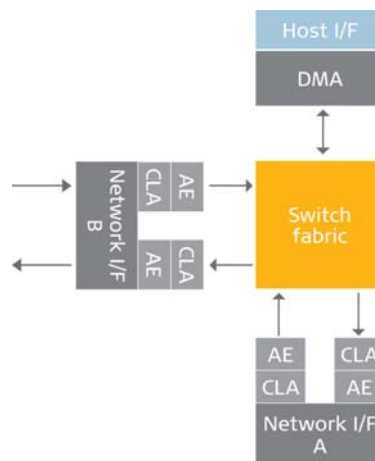


Figure 1: Vulcan Architecture (based on Commex patent-pending technology)

2.1 Vulcan Architecture – Benefits

Commex Vulcan architecture has numerous benefits including:

- Fewer packets sent to the processor - reducing processor utilization, allowing it to handle additional computational tasks, instead of classifying data that should be sent back to the network
- Packets sent to the correct core/processor - decreasing cache pollution per processor and enables improved server memory cache utilization
- Decreased amount of unnecessary traffic from Vulcan to the host processor - resulting in significant reduction in I/O power consumption from packet's round trip from the network to the processor and back
- Reduced unnecessary classification computing at the processor level – improving latency on computational tasks and decreasing processor power output
- Ability to cascade servers - each server handling only a specific portion of traffic
- Power savings - traffic is not sent through the I/O and memory buses

3 Implementation

This section details various application examples of Commex Vulcan and Commex patent-pending Content Aware Routing technology. For each example, the leveraged features, benefits and improved performance results are described.

3.1 Differential Services

The example is based on an automated trading application – a server resides in close proximity to the market exchange, received network data, and based on the data and sophisticated algorithms, automatically buys and sells stock. The buying and selling must happen with extremely low latency which remains constant. The server receives a very heavy volume of market data, in which there is a very small amount of data (e.g., 1% of the total) which is high priority. These are the “buy”, “sell” and “stop order” packets. The goal is to preserve the low latency of the high priority data while maintaining high throughput for the rest of the data which is lower priority (e.g., market historical).

To achieve the goal, there's a need to differentiate between the high and lower priority data in order to give higher priority to the high priority data. The high priority data is labeled in the payload with the text strings “buy”, “sell” and “stop order”. Without content awareness, it's not possible to perform this differentiation. Using Commex patent-pending Content Aware Routing technology, Commex Vulcan can look into the payload of the data packets and identify the target text strings. Then, the high priority packets can be sent to specific processing core(s) standing by just to process this small (e.g., 1% of total) amount of data. The rest of the lower priority data can be sent to the other processing cores. The result is that the high priority data can receive fixed low latency regardless of the total volume of data traffic, and, at the same time, high throughput can be attained for the lower priority data.

This is a complex task that demands the use of several mechanisms:

- Deep Packet Inspection (DPI) – searching for text that is not part of L2-L4 headers
- QoS for select packets
- Host interface per processing core – namely, DMA channel and interrupt vector

Based on the above requirements, most network adapters are unable to support such an application. Vulcan, however, with Commex Content Aware Routing technology, is able to classify the high priority packets, send them to specific processing cores and priority queues, and give priority to the customer's preferred and selected data set.

Figure 2 depicts latency for the high priority traffic as a function of load (volume for total data traffic). We compare Commex Vulcan with a leading conventional NIC. Commex Vulcan (the lower graph) shows more or less fixed low latency for the high priority traffic as load increases. The leading conventional NIC without Commex Content Aware routing shows increasing latency (by an order of magnitude) reaching an unacceptable level.

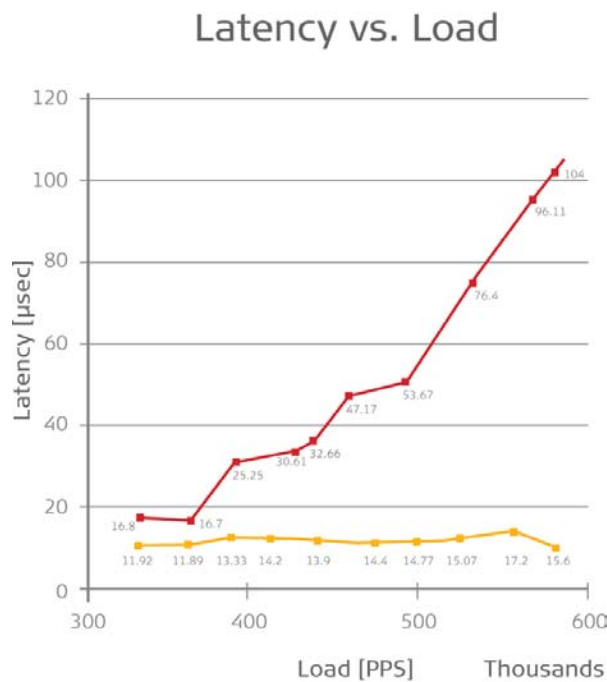


Figure 2: Latency vs. Load

3.2 Content Aware Bypass

In this example, there's an application which requires that only a particular type of packets (e.g., SCTP) is processed by the server. All other incoming packets sent to the server, should not be forwarded to the server host but rather sent back to the network,

or to other servers. Here too, most generic network adapters do not have the capability to directly send packets back to network, as they lack the ability to classify packets that are not IP, TCP or UDP.

Vulcan, with its structural configured classification, can be configured to detect the packet type and more importantly, can also distribute SCTP packets between all host cores. Vulcan directs the remainder of the traffic back to the network.

Figure 3, below details the data flow of an incoming packet as it enters Vulcan: (1) Traffic is classified; (2) SCTP packets are sent to the host processors; (3) Remaining traffic is sent to the other network port (4).

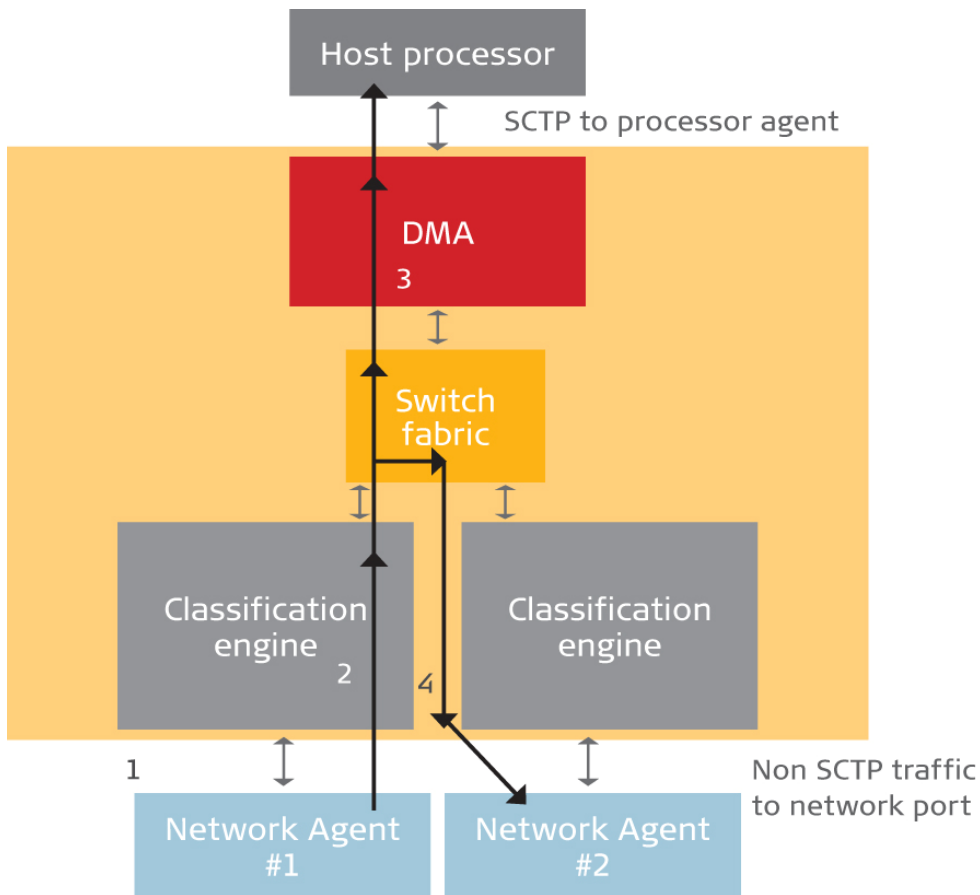


Figure 3: Content Aware Bypass

3.3 Network monitoring server Cluster

This example shows a network monitoring application for which the goal is to monitor all traffic; however, since traffic rate is 10Gbps, a single server with two processors is unable to process the total load. Accordingly, several servers can be arranged in a network monitoring cluster. The main requirement for traffic entering one server is that both the request and the response will be sent to the specific server and specific core at the server. Vulcan is able to classify the packet based on a 4-Tuple (IP source and destination added, L4 source and destination port), and either send the traffic to a

specific core within the server or to the next server in the cluster. The result is efficient load balancing between all cores in the cluster and persistency in data sent to the same processing core.

Figure 4 shows how: (1) The incoming traffic is classified; (2) A hash function is used and each packet is assigned a number. If the value is in the range defined by A (e.g. 0-15), the packet is distributed and sent to the server cores; (3) If the value is not in the range defined by A, the packet is sent to next server (4). This same action occurs in the second server for the range defined by B. Using these mechanisms enables the traffic to be distributed evenly between all servers and evenly between all cores of specific servers. As the classifier is a configurable machine, the number of servers in the cluster can be changed, enabling easy upgrades in process levels and traffic bandwidth.

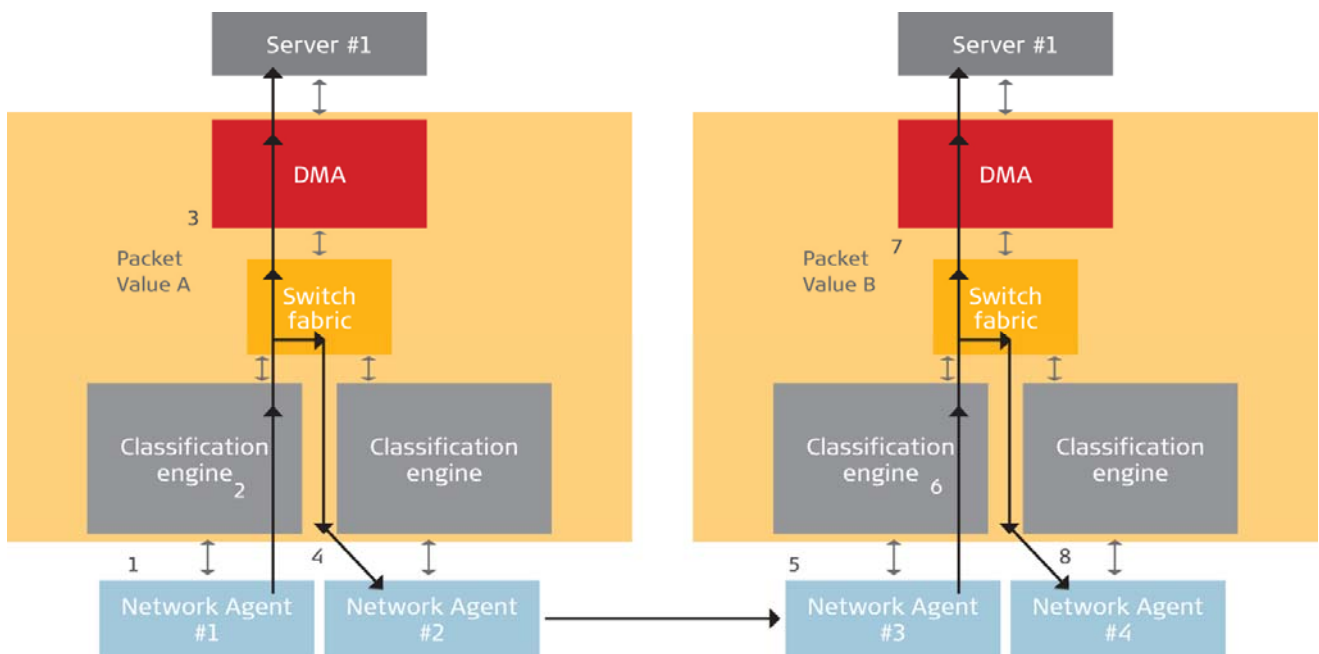


Figure 4: Monitor Cluster

4 Summary

Commex Vulcan NIC with patent-pending Content Aware Routing technology is a unique and innovative solution that seamlessly integrates switching and host interface concepts. It allows users to classify incoming traffic and direct it, whether partially or fully, to its intended destination server core or network port. Vulcan contains fast switch fabric that supports low latency and high throughput. Utilizing these mechanisms, Vulcan supports numerous applications, delivering higher bandwidth, lower latency, reduced computational load for network data traffic processing, as well as reduced cache pollution and power consumption. The overall result is more efficient utilization of the multiple processing cores on a multicore server.