Hindawi Publishing Corporation
International Journal of Reconfigurable Computing
Volume 2012, Article ID 342625, 14 pages
doi:10.1155/2012/342625

Research Article

A Dynamically Reconfigured Multi-FPGA Network Platform for High-Speed Malware Collection

Sascha Mühlbach[1] and Andreas Koch[2]

[1] *Secure Things Group, Center for Advanced Security Research Darmstadt, Mornewegstr. 32, 64293 Darmstadt, Germany*
[2] *Department of Computer Science, Embedded Systems and Applications Group, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany*

Correspondence should be addressed to Sascha Mühlbach, sascha.muehlbach@cased.de

Received 30 April 2011; Accepted 22 August 2011

Academic Editor: Marco D. Santambrogio

Malicious software has become a major threat to computer users on the Internet today. Security researchers need to gather and analyze large sample sets to develop effective countermeasures. The setting of honeypots, which emulate vulnerable applications, is one method to collect attack code. We have proposed a dedicated hardware architecture for honeypots which allows both high-speed operation at 10 Gb/s and beyond and offers a high resilience against attacks on the honeypot infrastructure itself. In this work, we refine the base NetStage architecture for better management and scalability. Using dynamic partial reconfiguration, we can now update the functionality of the honeypot during operation. To allow the operation of a larger number of vulnerability emulation handlers, the initial single-device architecture is extended to scalable multichip systems. We describe the technical aspects of these modifications and show results evaluating an implementation on a current quad-FPGA reconfigurable computing platform.

1. Introduction

The significant increase of malicious software (malware) in recent years (see [1]) requires security researchers to analyze an ever increasing amount of samples for developing effective prevention mechanisms. One method for collecting a large number of samples is the use of low-interaction honeypots (e.g., [2]). Such dedicated computer systems emulate vulnerabilities in applications and are connected directly to the Internet, spanning large IP address spaces to attract many different attackers. A number of software applications are available helping in building up honeypot systems. But in addition to having performance limitations in high-speed environments (10+ Gb/s), such software systems also suffer from being compromisable themselves (they can be subverted to attack even more hosts). Given the experience of the Nepenthes research project [3], it is extremely hard to realize an attack surface of millions of IP addresses (such as multiple class B networks) with actively communicating service modules running in software on a single server.

In this context, we have proposed MalCoBox, a low-interaction malware-collection honeypot realized entirely in reconfigurable hardware without any software components in [4]. The core of the MalCoBox system is NetStage, a high-speed implementation of the basic Internet communication protocols, attached to several independent vulnerability emulation handlers (VEH), each emulating a specific security flaw of an application (see Figure 1). We have demonstrated the feasibility of that approach by implementing a prototype on a FPGA platform, fully employing the power of dedicated hardware resources to support 10+ Gb/s network traffic.

With NetStage able to sustain above wire-speed throughput, MalCoBox can easily handle the emulation of multiple class B networks in a single system (Honeynet-in-a-Box), thus avoiding the overhead of administering a honeynet distributed across multiple servers. Beyond the performance aspects, in context of the network security domain, a purely hardware-based approach such as ours has the additional advantage that no general-purpose software programmable

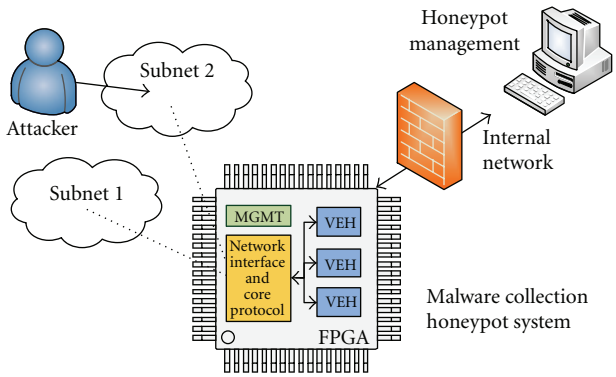

Figure 1: Hardware-based malware collection.

processor is present that could be subverted if the honeypot itself is being attacked.

An important issue for potential MalCoBox users is how the platform can be updated during operation with new or improved vulnerability emulation handlers (VEHs) to react to the changing threat landscape. For an FPGA-based system, the hardware functionality can be altered during operation by using partial reconfiguration (PR). This approach has already been used for network routers in [5]. We now employ the technique in a larger scope to flexibly swap-in new VEHs while the rest of the system stays in operation. The initial discussion presented in [6] is expanded in this work.

Another aspect of great practical interest is the number of different vulnerabilities that can be emulated in *parallel*. The original MalCoBox relied on a single-device implementation of NetStage and was limited to ca. 20 VEHs active in the system. This is a gap to software honeypots, where even the low-interaction variants often support $50 \cdots 100$ different vulnerabilities implemented as scripts in languages such as Perl and Python. While it could be argued that the capacities of individual FPGA chips does increase from each generation to the next (which they do), the larger high-end devices are significantly more expensive per logic cell than the mid-range versions. Thus, it is worthwhile to examine how the MalCoBox capacity can be extended using a multidevice NetStage implementation. This approach has been introduced in [7] and is described in greater detail here.

The paper is organized as follows: Section 2 briefly describes the core architecture components. Section 3 covers details of the ring implementation and elaborates the differences between single-chip and multichip solution. Section 4 continues with a description on the required modifications of the partial reconfiguration strategy. The implementation of the complete system on the BEEcube BEE3 quad-FPGA reconfigurable computing platform [8] is described in Section 5, followed by experimental results given in Section 6. We close with a conclusion and an outlook towards further research in the last Section.

1.1. Related Work. To our knowledge, this is the first implementation of such a honeypot system using pure dedicated hardware blocks. In 2007, Pejović et al [9]. presented an initial concept for a hardware honeypot with RAM-based state machines for the emulations. Unfortunately, they did not give any detailed results on the achievable performance and the possible parallelism. It is likely, however, that the RAM-based state machines could become a bottleneck in high-speed environments due to limited bandwidth. Thus, our architecture contains only dedicated hardware blocks.

In terms of FPGA-based networking, a popular generic platform for research is the Stanford NetFPGA [10], containing an FPGA, four 1G interfaces, and various memories. The platform is the vehicle for a wide spectrum of research, for example, on accelerated switches, routers, and network monitoring devices. Internally, NetFPGA provides a flexible data-path structure into which custom processing modules can be easily inserted. With the widespread use of NetFPGA, a new version supporting 10 G networks is currently being released.

Another related research project is DynaCORE [11]. It consists of a network-on-chip- (NoC-) oriented architecture for a generic reconfigurable network coprocessor, combining general network processing in software with accelerated functions (e.g., encryption) in hardware units. By using current techniques such as partial reconfiguration, the platform can be adapted to different communication situations.

In [12], the authors present a reconfigurable network processor based on a MicroBlaze softcore-CPU extended with multiple exchangeable hardware accelerators. Partial reconfiguration is used to adapt the system to the current network traffic workload. Partial reconfiguration is also employed in [13] to support network virtualization, realizing multiple virtual routers in hardware on a single FPGA.

In contrast to these often packet-oriented approaches, our own research has always been aiming at higher-level Internet (e.g., TCP and UDP) and application protocols. We, thus, have created NetStage [4] as a novel base architecture for our honeypot appliance. NetStage is built around an all-hardware Internet protocol stack implementation (including TCP operation). We have chosen to follow some of the approaches proven useful with NetFPGA, for example, the capability to insert "plug-in" hardware modules at various points in the processing flow. In contrast to DynaCORE, however, we generate a light-weight application specific interconnect between these modules, instead of using a general-purpose, but larger NoC scheme.

2. Key Architecture Components

Figure 2 shows the base NetStage architecture (discussed in greater detail in [4]), including extensions to support dynamic partial reconfiguration (DPR). The architecture provides module slots (Figure 2(a)) into which the partial VEH bitstreams can be loaded. These VEH slots are loosely interconnected with the core system by buffers, allowing all VEHs to have the same external interface (important for DPR). Thus, any VEH may be configured into any of the slots of the same size, with the buffers limiting the impact of brief VEH-level stalls on the system-level throughput.

VEHs share the underlying implementations of the core protocols (IP, TCP, and UDP) in NetStage. These have

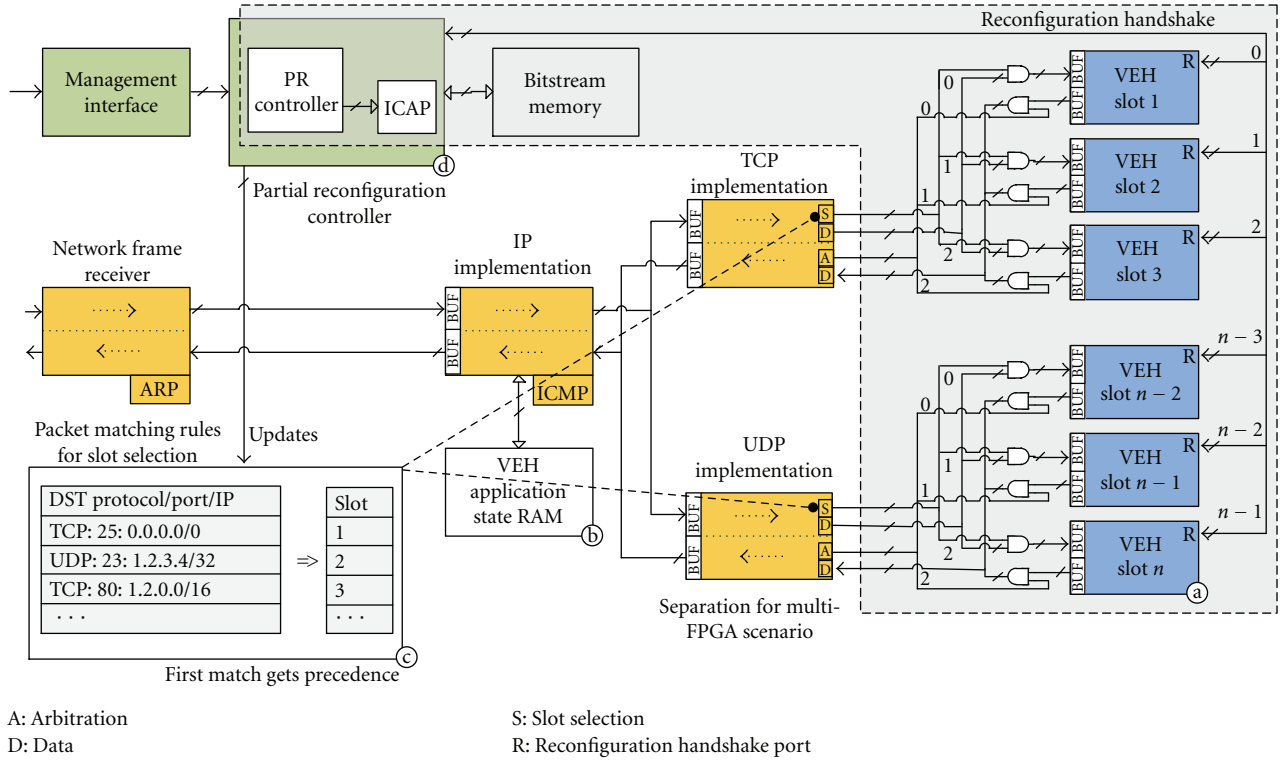

Figure 2: Core architecture of the partially reconfigurable malware collection system.

been very carefully optimized to achieve a throughput of at least 10 Gb/s by using pipeline- and task-level parallelism to keep up with the line rate of the 10 Gb/s external network interface.

In some cases, VEHs have to track session state to generate an appropriate response. NetStage provides a central facility for storing per-connection state (Figure 2(b)). When a packet is passing the IP implementation, the globally maintained state information is attached to the packet in a custom control header which accompanies every packet through the system. The VEH can read this information, act on it, and update the value if necessary. The modified header is written back to the state memory when a response packet (or an empty state write packet) passes the IP implementation on the transmit path. Such a centralized storage is more efficient than attempting to store state in each VEH (which would fragment the capacity of the on-chip memories).

The global VEH application state memory can also be used to save/restore VEH state during reconfiguration to allow the seamless swapping-in of newer (but state-compatible) versions of a VEH.

2.1. Vulnerability Emulation Handler. When a packet has passed through the NetStage core, it will be forwarded to the responsible slot, where the VEH performs the actual malware detection and extraction. Packets are routed to the appropriate slots by means of a routing table (Figure 2(c)) that holds matching rules for the different vulnerability emulations currently active in the system. The table is writable to allow dynamic modification of the actual VEHs used. A basic set of matching rules includes the destination port, destination IP and netmask. The latter allows us to set up separate IP address ranges which use VEHs for different vulnerabilities on the same port (e.g., many handlers will listen on the HTTP port 80).

With the processing speed achievable using reconfigurable hardware, these basic rules could also be extended to directly match packet contents instead of just header fields. Dedicated matching units could search the payloads of packets being forwarded through the NetStage layers and have the matching results available in time for the slot routing decision. However, this would require dynamic reconfiguration of the actual matching units together with the VEHs when they change instead of just writing new values into registers (as in the basic header-only matcher). Since all our current VEHs are selected just based on protocol and port (independently of the payload), we can continue to use the simpler basic approach.

2.2. Management Section. The partial dynamic reconfiguration of VEHs is managed by the partial reconfiguration controller (PRC, Figure 2(d)), which is connected to the FPGAs internal configuration access port (ICAP). On the application side, the PRC is connected to the MalCoBox management interface (either by a PCI express endpoint or a dedicated network link, depending on the selected deployment mode of the system). The PRC is also connected to the individual VEH slots by a number of handshake signals to inform the VEHs about their impending reconfiguration

(for a clean shutdown, etc.) and to check whether the VEH is idle. An attached bitstream memory can hold several partial bitstreams to allow the system to be reconfigured independently of the management station in future implementations.

2.3. Reconfigurable VEHs. The VEHs are dedicated to emulate known application vulnerabilities that may be exploited by malware. As malware is evolving quickly and new vulnerabilities are detected on a regular basis, the VEHs need to be updated as well. For example, when the system is used for research purposes, multiple updates per day would commonly be required if a researcher would like to react to current observations in the network traffic.

In a conventional software-based system, such changes could generally be handled quickly by altering a few lines of code in a script language. However, when using dedicate hardware, even minor changes of the functionality require full logic synthesis, mapping, and place-and-route steps. This effort can be significantly reduced by employing dynamic partial reconfiguration (DPR) to recompile just those VEHs actually affected by the change.

As a secondary effect, DPR also allows the arbitrary combination of VEHs in the system. Otherwise, each specific set of selected VEHs would need to be compiled into its own static configuration.

To support independent partial reconfiguration of any of the VEH slots within the architecture, a wrapper encapsulates the actual VEH implementation module (see Figure 3). This wrapper includes glue logic controlled by the partial reconfiguration controller to disconnect/reconnect all inputs and outputs of the VEH module. This clean separation is essential to avoid introducing errors in the rest of the system when reconfiguring.

The wrapper also contains the send and receive buffers for each module as well as the corresponding buffer management logic. As all the handlers share the same buffer structure, it is more efficient to keep it static than configure it with each VEH. The inputs and outputs of the wrapper are directly connected to the MalCoBox core (see Figure 2).

3. Multidevice Architecture

To extend our system to multiple FPGAs, we will draw the boundaries between the static NetStage core (basic Ethernet and Internet protocol functions) and the dynamically exchangeable VEH slots (see the dotted line in Figure 2). One FPGA acting as Master node holds the network core and the network interfaces, and the remaining other FPGAs, called VEH nodes, contain the individual emulation blocks. The BEE3 platform supports a number of interdevice interconnection schemes. For future scalability independently of the BEE3 architecture, we decided to implement a ring structure. Such rings have already proven useful in multichip systems internally using NoCs [14].

For extending the NetStage-based MalCoBox to multiple devices, a unidirectional ring suffices (Figure 4). The unidirectional ring needs fewer I/O pins on the FPGAs and avoids the increased latency of the bus-turnaround cycles that would be required when running a bidirectional bus over the same pins. Also note that in contrast to the implementations in discussed in prior work, the one described here is using the external SDRAM instead of the internal BlockRAM to temporarily store bitstreams, conserving FPGA resources to allow more VEHs.

3.1. Ring Communication. For the communication on the ring, we use 66 of the 72 available inter-FPGA data lines on the BEE3 which are run in DDR mode, resulting in 132 bits of data per clock cycle [7]. Four bits are reserved for status bits, the remaining 128 form the data transmission word. As data words are sent continuously on the ring for synchronization purposes (even if not actual data needs to be transmitted), a valid flag is used to indicate words holding actual message data. For the separation of individual messages, we use two flags to signal the first and the last word of a message. As the actual byte size of the message is already stored within the internal control header (ICH, see Figure 8) used by NetStage (prefixed to the message body), no special-case processing is required for unused bytes in the last data word of a message. A final flag is used to denote special ring control messages used, for example, to control the DPR process (see next section). In a future extension, this could also be used, for example, to enumerate the ring nodes automatically during initialization. Currently, the destination addresses of the available FPGAs inside the ring are set during compile time.

The ICH-prefixed message is prefixed yet again with a 128 b ring control header (RCH) when it is transmitted between devices. The RCH carries the type information of the message and the destination FPGA. The remaining bits are reserved to implement further control functions in the future.

Since we want to maintain a high bandwidth and low latency even when distributing the architecture across multiple devices, we want to operate the interdevice links at maximum speed. To this end, we use the BEE3-provided global clock to all FPGAs as the ring clock. However, due to different trace lengths and other board-level signal integrity effects, reliable operation at our target frequency of 250 MHz requires a training of the individual FPGAs to the link characteristics. This is done using the technique proposed in [15]. A known training sequence is transmitted between adjacent FPGAs, and the receiver adjusts its delay until it receives a stable pattern from the transmitter.

Two additional signals are required to realize training procedure (see Figures 5 and 6). The Master starts the process by asserting a Sync signal, which is routed around the entire ring. It is used to both initiate training between neighboring FPGAs and to test whether the nodes did configure correctly on start-up (an error is indicated if the Sync sent out by the Master does not match the incoming Sync passed around the ring). Once the receiving FPGA of a synchronization pair has locked on to the training pattern, it asserts its internal Ready signal, which is ANDed with the Ready incoming from its transmitting partner before being output to its receiving partner (the next device on the ring). Once the Master has received an asserted Ready signal passed around the entire

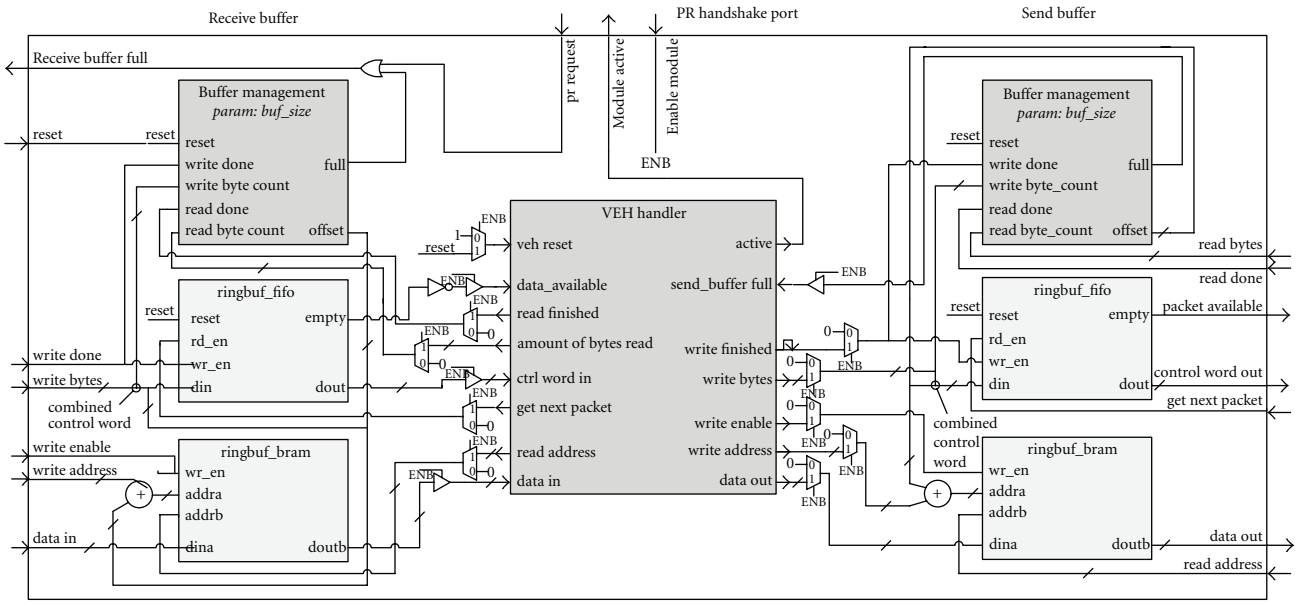

Figure 3: Wrapper encapsulating a vulnerability emulation handler slot.

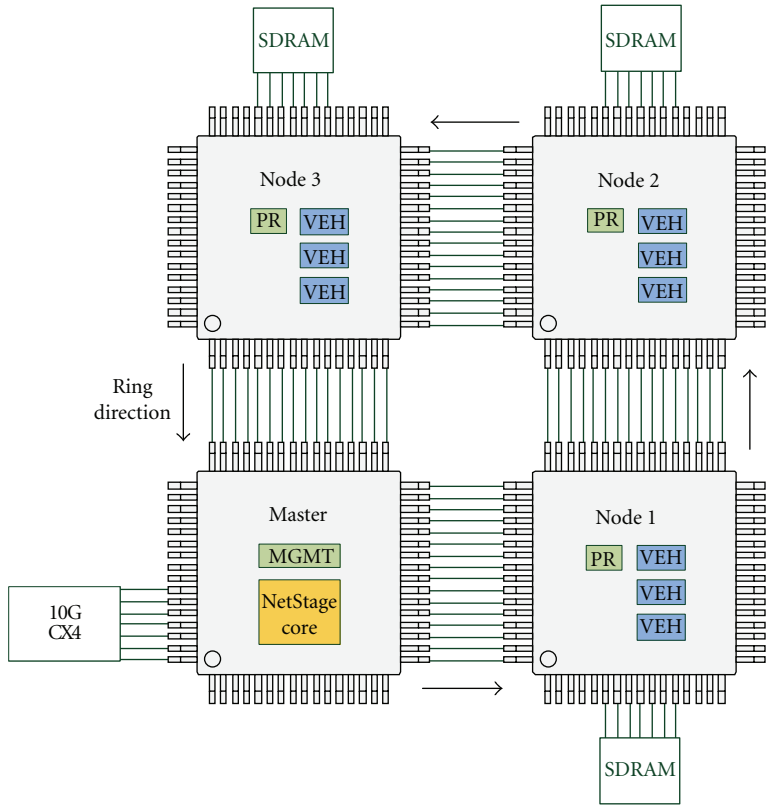

Figure 4: Multi-FPGA network processor in ring topology.

ring, it stops training and releases the ring into normal operation.

The ring thus achieves a transfer rate of 32 Gb/s between nodes, more than sufficient for our current 10 Gb/s network environment. For simplicity, and since we did not experience any data integrity issues in our practical experiments once training completed, we do not perform error detection/correction on the ring communications. However, for long-term production use, CRC/ECC facilities could be added here. As there are still data lines available on the BEE3, this could be easily implemented without affecting the base architecture.

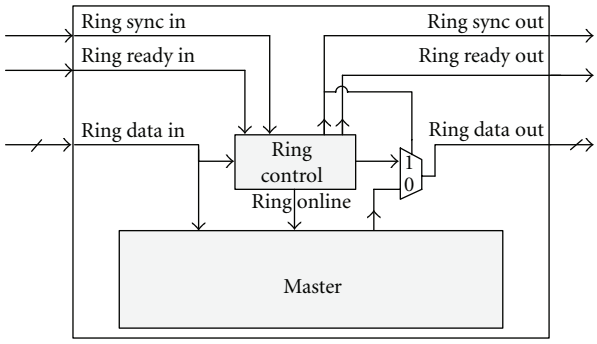

Figure 5: Schematic overview of the master.

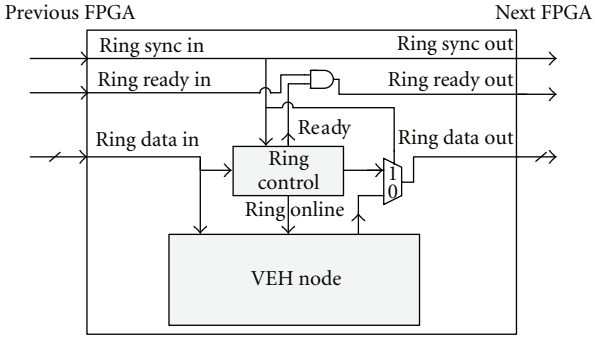

Figure 6: Schematic overview of the nodes.

3.2. Master Node. Beyond the network core and the management section that was already present in the single-chip NetStage implementation, the Master node (see Figure 7) now contains additional logic (Figure 7(a)) to handle the ring communication. In particular, this includes send and receive interface modules as well as the FPGA addressing logic. Note that we do not place any VEHs in the Master node and the currently unused space is intended to be used for future extensions of the NetStage core (e.g., to IPv6). Thus, the Master itself will not be dynamically reconfigured and does not require an internal ICAP controller. However, the Master is responsible for initiating the reconfiguration of the VEH nodes. Thus, the management section in the Master and the configuration controllers in the VEH nodes interact, which is achieved by specialized ring control messages.

The payload data traffic around the ring is organized on two levels. The 32B ICH (see Figure 8) replaces the original protocol headers for a packet with a more compact representation. It also carries the packet-to-handler routing information of a message on the ring in the form of a destination VEH node ID and the VEH slot on that node. Since the destination node ID is already specified in the RCH, this might be seen as redundant. However, the RCH is present only while a packet is transmitted between nodes and stripped from it for node-internal processing. Since we want to give VEHs the ability to transparently forward packets to other VEHs which might also reside on other nodes, they can read/write that destination data in the ICH instead (which, due to alignment reasons, is not efficient for performing ring-level routing).

As we now have multiple destination FPGAs, the Master routing table, which associated packets just with the responsible VEH slot in the single-chip version, needs be extended to hold the destination node ID as well (Figure 7(b)). This is used to build the RCH when the packet is sent out over the ring. To reduce the latency, the process to look up the destination address (Figure 7(c)) is pipelined between the core and the Ring Send module. This can be easily done as packets are not reordered between the two modules.

The Master will silently discard packets not matching any rule in its routing table to conserve bandwidth on the ring links. Furthermore, core IP protocols such as ARP and ICMP are usually handled with low-latency entirely inside the Master, and do not cause ring traffic, either.

3.3. VEH Node. The individual VEHs are housed in the VEH nodes (see Figure 9). For communication with the rest of the system, the VEH nodes need the same ring interface modules (Figure 9(a)) as the Master node. Furthermore, a node-local packet distributor and aggregator (Figure 9(b)) emulate the single-chip NetStage core interface so that VEHs can be attached directly connected to the network core of the single-chip implementation. VEHs are thus portable between the single- and multichip versions.

In contrast to the Master node, the VEH are actually dynamically reconfigured to exchange VEHs. Thus, they do need a PR controller and access to the ICAP. The details of this are described in greater detail in the next section.

The ring receive module in each VEH node checks the type of an incoming ring message and its destination address field and either forwards the packet to the local distributor module, a reconfiguration message to the local PR controller, or immediately inserts the message into the forwarding queue if is intended for another node. VEH response network packets are picked up by the packet aggregator and inserted into an output queue, passing it on around the ring until it reaches the external network connection at the Master node.

4. Partial Reconfiguration

Partial bitstream data is transferred from the management station (usually an external PC or server) to the MalCoBox via the management interface. As the MalCoBox should be able to run as an appliance under remote management, we implemented a stand-alone on-chip reconfiguration interface instead of using the JTAG port together with a software programmer on a host PC.

The underlying protocol used to transmit the bitstream to the MalCoBox consists of the raw bitstream prefixed by a reconfiguration header (see Figure 10). The header contains the bitstream size, the FPGA slot location information, and the rules for the Master node routing table to direct packets to the newly configure VEH.

In contrast to the single-chip implementation [6], in the multidevice scenario, the management interface housed in the Master does not have a direct connection to the partial reconfiguration controller (PRC). Instead, the bitstream is transferred over the ring to the destination VEH node

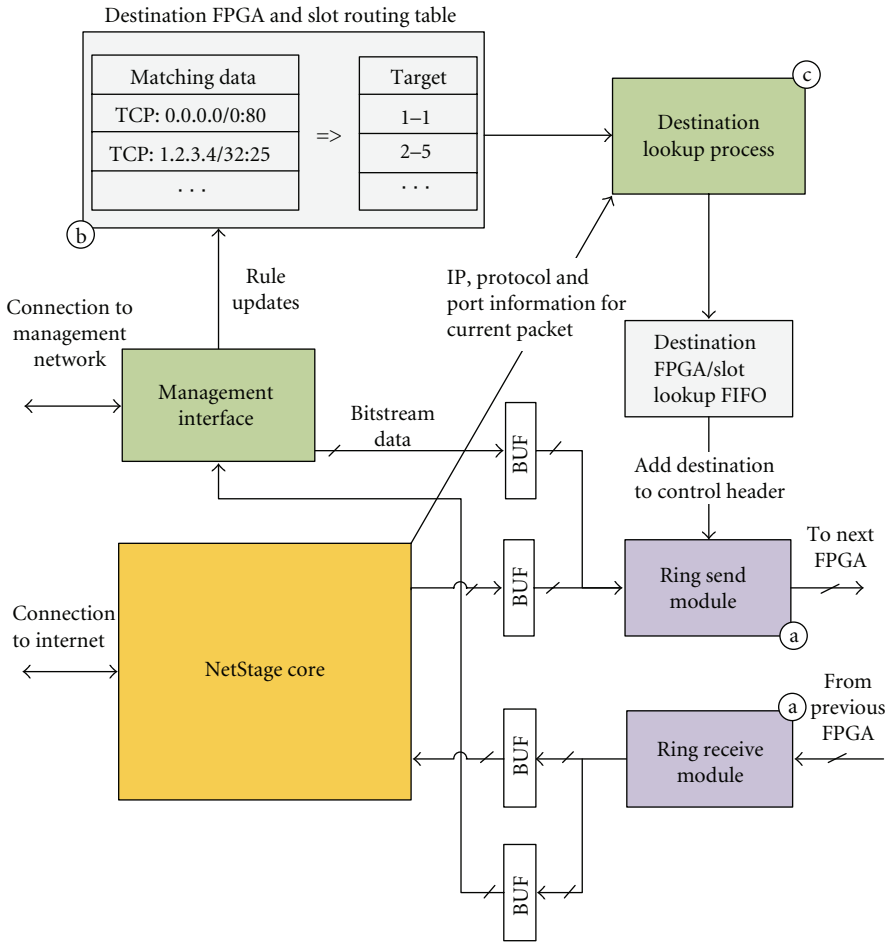

Figure 7: Master node architecture.

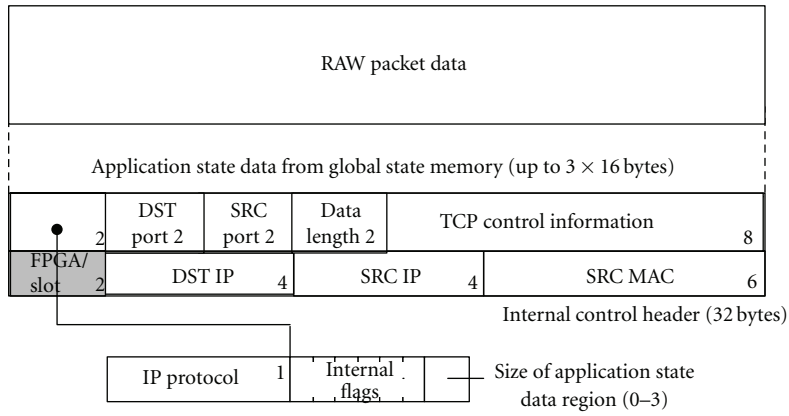

Figure 8: Structure of the internal control header.

FPGA, but the routing rules table still remains inside the Master node. The management interface, therefore, extracts the header information from incoming reconfiguration data requests and updates the routing table, while the raw bitstream data is forwarded to the FPGA specified in the reconfiguration header (see also Figure 7). As the partial reconfiguration process is now distributed across multiple devices, the time between the Master-local routing rule update and the activation of new VEH in a remote node is longer than on the single-chip system. Thus, to avoid misrouting of packets, the new routing rule is explicitly disabled until the VEH is actually ready to accept traffic.

In a VEH node, an incoming bitstream is stored in node-local external DDR-SDRAM memory. One an actual

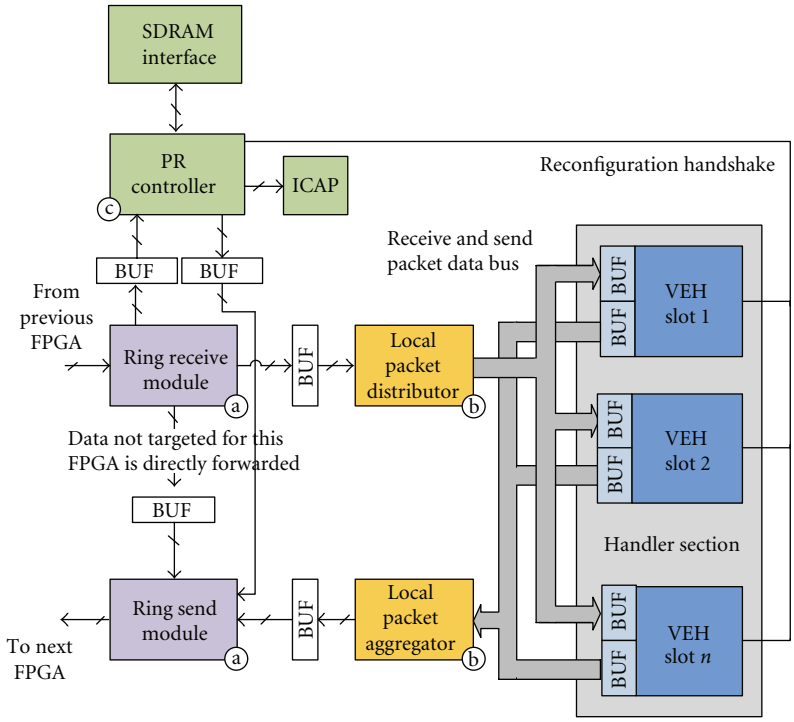

Figure 9: VEH node architecture.

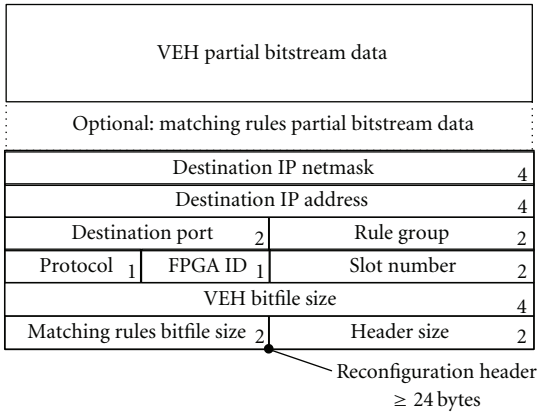

Figure 10: Custom PR header and bitstream data.

reconfiguration is requested, a fast DMA unit retrieves the bitstream data from memory and transfers it at maximum speed to the ICAP configuration interface. This two-step approach could also be used in a later extension to, for example, integrity-check the bitstream for communication errors, or to accept only signed bitstreams [16, 17]. Since the ring communication has proven reliable in our tests, and the management console is trusted, the current prototype does not implement these facilities.

4.1. Partial Reconfiguration Process. The distributed reconfiguration process is performed in the following order.

(1) The rule header of incoming bitstream data is extracted, and the rule table is updated with the new rule (eventually replacing an existing one), having the active flag set to zero.

(2) Incoming bitstream data is forwarded to the corresponding VEH node.

(3) After complete reception of bitstream data, the PRC in the VEH node starts the reconfiguration process.

(4) After completion of reconfiguration, the PRC sends a DONE status to the Master as a ring control message.

(5) The Master management interface receives the message and activates the routing rule so that packets will actually be forwarded.

Network packets and reconfiguration messages (including the bitstream data) share the ring. However, since reconfiguration management is crucial for the reliable operation of the system, these ring control messages receive priority over regular packet transmissions.

Internally, the reconfiguration process in the VEH nodes follows the approach implemented for the single-chip solution [6]. When the node-local PRC receives a reconfiguration request, it initially informs the wrapper of the target slot that the slot is about to be reconfigured. This will stop the receive buffer of the VEH from accepting new packets. The VEH is allowed to process all of the packets held in the buffer at this time, asserting a signal to the PRC on completion. The PRC then deactivates the VEH, and the now inactive VEH is disconnected from the slot wrapper. The actual bitstream data is then read from the DDR-SDRAM and fed into the ICAP. Once reconfiguration is completed, the PRC re-enables the VEH-wrapper connections and allows the new VEH to wake up in its reset state.

5. Implementation

The MalCoBox running on the multidevice NetStage architecture has been implemented on the BEEcube BEE3 FPGA-based reconfigurable computing platform, which is equipped with eight 10 Gb/s network interfaces and four Xilinx Virtex 5 FPGAs (2x LX155T, 2x LX95T). The Master node is realized as one of the smaller SX95Ts to have both of the larger LX155T devices available for VEHs.

Network connectivity is provided by the Xilinx XAUI and 10 G MAC IPs. The network core in the Master runs at the speed of the 156.25 MHz clock of the 10 G network interface. Together with the internal bus width of 128 bit, this leads to a maximum core throughput of 20 Gb/s. This overprovisioning allows us to react to brief stalls in the data flow. These could occur if complex VEHs needed extra time to process a packet (e.g., perform a DRAM lookup) and could not guarantee a steady 10 Gb/s throughput. The 20 Gb/s throughput supported by NetStage thus allows handlers to "catch up" with the normal 10 Gb/s traffic by burst-processing the data accumulated in the buffers at double speed.

To allow for greater design flexibility and allow the execution of more complex VEHs having longer combinational paths, the clock rates of the Master and VEH Nodes can be set independently. Currently, we run the VEH Nodes at 125 MHz, thus supporting burst processing at a rate of 16 Gb/s. This can be easily altered (sped up or slowed down) to match the complexity (delay) of the required VEHs, as long as the 10 Gb/s minimum throughput is always achieved.

The ICAP is operated at 32 b data width and driven by a separate clock to support variable reconfiguration speeds (and thus support experiments with overclocking the ICAP). Management access is implemented as dedicated network interface with a unique MAC address, directly connected to a standard desktop PC or server. The management interface receives bitstream data and control operations over the network using a custom protocol. Perl scripts are used to assemble the appropriate network packets. The DDR2-SDRAM interface in the VEH nodes is realized by a Xilinx MIG core and fully uses the DDR2-SDRAM bandwidth.

The size of all intermodule and slot buffers is set to 4 kB (to hold 2 packets with a maximum size of 1500 B), which is sufficient to assure stall-free operation as the modules generally consume the data at a minimum rate of 10 Gb/s). If many variable-latency VEHs requiring burst processing are employed, the buffer size will need to be increased correspondingly. This has already been done for the ring receive buffers, which are set to 16 kB to provide sufficient headroom to receive bursts of packets on the ring. The size of the global application state memory in the Master node is currently set to 1 Mb of BRAM, which is sufficient to manage the short running sessions (only a few seconds each) that we expect in the honeypot use case. By session, we mean the time interval the attacker needs to check whether the target is vulnerable until the reception of the attack code. For applications requiring more state storage, the data could also be stored in the per-node DDR2-SDRAM available on the BEE3. This would incur longer access latencies, though. These could be avoided on newer hardware platforms by using low-latency external memory technologies such as Bandwidth Engine [18] or Hybrid Memory Cube [19].

5.1. Destination Lookup. The rules that control the message routing to VEHs in different nodes are stored in the destination routing table (see Figure 11) inside the Master node. This table is implemented as BRAM (currently with a size of 1024 rules), to achieve high lookup speeds and flexible scaling. In addition to the routing information, each rule entry contains a rule ID that is used for management purposes and an active flag used during the reconfiguration process. The table supports multiple rules for the same destination VEH (e.g., to let it respond to different IP addresses).

As the destination routing decision is on the critical path with regard to latency, we use a hierarchical approach for lookups. Rules with the same destination port are represented as a linked list, and a CAM is used to retrieve the BRAM address of the list head for a given port (see Figure 12). Then, the individual rules for this port are searched in list order by following the rules' next pointers. Beyond quick lookups, this also ensure that rules with the longest IP address prefix will be matched first. The management process ensures that rules are inserted in the correct order.

For efficiency, we have restricted the CAM size to 256 entries, reasoning that 256 different active ports should be sufficient for most cases. Since the CAM has an 8b wide output, the heads of the per-port rule lists always start in the bottom 256 addresses of the routing table BRAM.

5.2. Example VEHs. To test the system, we have created a number of VEHs emulating different vulnerabilities and applications. In addition to controlling FSMs, the VEHs contain additional logic to perform tasks such as fast parallel pattern matching.

(1) *SIP*. The SIP VEH looks for packets exploiting a vulnerability of the software SIP SDK sipXtapi [20]. The exploit uses a buffer overflow occurring if a SIP INVITE packet contains a CSeq field value exceeding 24 bytes in length. This VEH is based on the UDP protocol.

(2) *MSSQL.* Another UDP-based VEH has a similar structure and is emulating a vulnerable MSSQL 2000 server looking for exploits targeting the resolution service [21]. This exploit was used in the past by, for example, the Slammer worm.

(3) *Web Server.* As a VEH for a further popular application, we implemented a simple web server emulation that contains a ROM with predefined HTML pages to be served to clients. The HTTP headers needed for response generation are also stored inside the ROM. An FSM checks the URL of incoming requests and fetches the corresponding output data to be sent from the ROM. This VEH can be flexibly used, for example, to emulate a login page for a company intranet and monitor

Target FPGA and slot selection table (BRAM)									
Addr.	Protocol	Target FPGA	Target Slot	Port	Netmask	IP addr.	Rule ID	Rule Act.	Next Rule
(10 b)	(8 b)	(8 b)	(8 b)	(16 b)	(32 b)	(32 b)	(16 b)	(1 b)	(10 b)
0	0x06	1	0	80	0x00000000	0x00000000	13	0	0
1	0x06	1	1	25	0xFFFFFFFF	0x53251021	25	1	256
...	...	...	...	...	...	...	...	...	...
256	0x11	2	3	25	0xFFFFFF00	0x10102500	47	1	257
257	0x06	1	1	25	0xFFFFFF00	0x32122500	69	1	0
...	...	...	...	...	...	...	...	...	...

⟶ Linked list to speed up lookups and to maintain IP/netmask prefix order for matching

Figure 11: Layout of the destination lookup table.

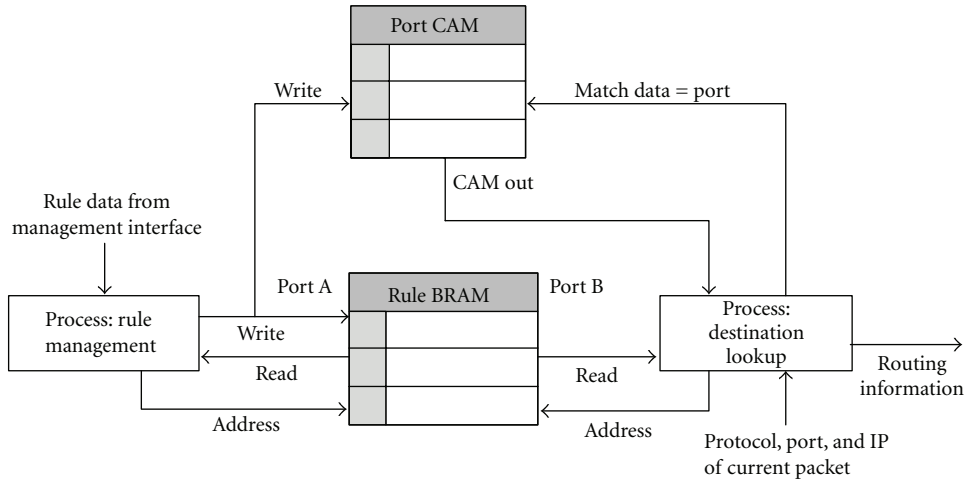

Figure 12: Implementation of the destination lookup process.

attack attempts (e.g., brute force logins) or attacks to the web server itself.

(4) *Mail Server.* As spam is amongst the widespread distribution techniques for malware, we implemented a mail server VEH that accepts incoming mails and pretends to be an open relay server. It contains a FSM that implements the basic SMTP dialog for the reception of mails.

6. Results

The design was synthesized and mapped using Xilinx ISE 12.4, targeting a SX95T for the Master node and both SX95T and the LX155T devices as VEH nodes. Partial reconfiguration was implemented using the latest partial reconfiguration flow available in PlanAhead 12.4 [22]. Each VEH node was configured to include 24 slots. The VEH module slots were placed manually on the FPGA and sized based on the resource usage trends shown by the sample VEH synthesis results. The resulting layout can be seen in Figure 13. To support VEHs with different resource needs (BRAM vs. LUTs), four kinds of slots, differing in the number and types of contained resources (see Table 5), are provided.

As techniques to dynamically relocate bitstreams on the FPGA matrix are not yet production ready, and even research versions have significant limitations (e.g., only support for outdated device families), we have to create separate bitstreams for the all of the different slots a VEH can be placed. In addition to requiring more storage, this also necessitates to run the place-and-route tools multiple times with different area constraints. Each run produces the partial bitfile for a specific VEH-slot combination. However, due to using partial reconfiguration, bitfile sets for different VEHs can be created and used independently, for example, exploiting a multicore server by executing many tool runs in parallel.

System tests were performed by simulation as well as on an actual BEE3 machine connected to a quad-XEON Linux server, sending data to the VEHs at 10 Gb/s. Partial reconfiguration was performed under operator control, loading in new bitstreams via network from the management station.

6.1. Synthesis Results. The synthesis results for all components are given in Tables 1 and 2. For the VEH nodes, we show results only for the LX155T, as the results for the SX95T are very similar (in terms of resource requirements).

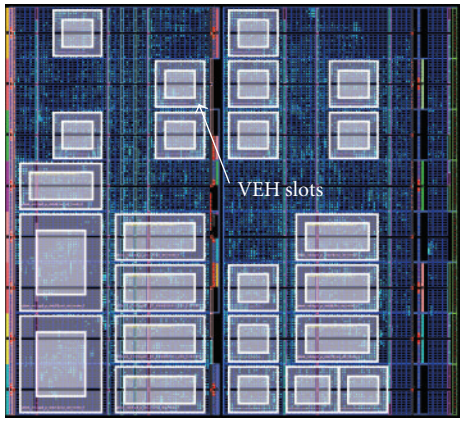

Figure 13: FPGA layout for 24 VEH slots.

The NetStage core on the SX95T Master node requires around 20% of the LUTs and 38% of the BRAMs. The high number of BRAMs is due to several buffers and the global application state memory. The mapped design including IP blocks occupies around 28% of the LUT and 47% of the BRAM resources distributed amongst 50% of the slices. This still leaves sufficient area unoccupied on the SX95T to allow for further extension of the Master node functionality.

The number of available BRAMs is crucial for our platform (due to the multiple buffers). As the SX95T and the LX155T have nearly the same number of BRAMs, these mapping results confirm our decision to put the Master node into the SX95T and leave the large number of LUTs inside the LX155T available for VEHs.

In the VEH node, the ring interface, the partial reconfiguration controller, and the VEH slot interfaces occupy around 20% of the FPGA. This leaves nearly 80% of the LUT resources available for the actual VEH implementations. In practice, the total number of slots per FPGA is limited by the number of BRAMs available to implement the slot buffers (five BRAMs are needed per slot). From these results, we conclude that we could theoretically support up to 36 VEH slots per FPGA on both the LX155T and SX95T.

In comparison to our single-chip implementation, which could hold 20 VEHs together with the NetStage core on a single LX155T device, the multi-FPGA approach is a significant improvement of the total processing power of our platform. When using all three VEH node FPGAs to their full extent, the system supports the parallel operation of 100 VEHs (depending on module size), which should suffice even for very complex honeypot use cases.

6.2. VEHs. Table 3 summarizes the area requirements for the various VEH modules. They are only showing little variation, which is advantageous for putting them into different slots on the FPGA. Amongst them, the SIP VEH requires the most LUTs, as it contains the most complex pattern matching algorithm. Overall, the VEHs are relatively small compared to the device capacity, thus we are confident that our slot numbers are realistic.

Table 1: Synthesis results for Master node components.

Module	LUT	Reg. bits	BRAM
Network core incl. management	12,297	8,884	93
Ring interface	788	1,489	16
Mapped incl. MAC, XAUI and clocks	16,532	13,526	117
In % of SX95T	28	22	47

Table 2: Synthesis results for VEH node components.

Module	LUT	Reg. bits	BRAM
Ring interface	976	2,048	20
PR controller	722	544	4
VEH section with 24 slots (w/o VEHs)	15,494	6,426	120
Total incl. MIG, without VEHs	19,540	12,428	150
In % of LX155T	20	12	70

Table 3: Synthesis results for the VEHs.

Module	LUT	Reg. bits
SIP VEH	1082	358
MSSQL VEH	875	562
Web server VEH	1026	586
Mail server VEH	741	362

6.3. Performance. The actual response time depends on the latency of the platform and the speed of the VEHs. As these numbers are, in turn, highly dependent on the implemented functionality, and the distribution of incoming network traffic, we show numbers for the upper and lower limits. For these experiments, we consider different fill levels of the buffers inside the NetStage core and the ring: all buffers empty (the best case), nearly half full (average case), and nearly full (worst case). For simplicity, we assume that all buffers in the system have the same fill level and that the VEHs are able to actually sustain a speed of 10 Gb/s (possible using the sample VEHs described above).

Table 4 lists the total round-trip times (RTTs) for a 1000 byte request packet that generates a 1000 byte response packet. As the packets have the same size, the time is independent of the ring location of the device holding the measured VEH.

Obviously, the fill level of the buffers inside the ring nodes has a severe impact on the latency, inducing a 10x increase in latency between empty and nearly full buffers. However, as we are currently feeding the system with only one 10 G interface, and the VEHs are all designed for high-speed operation, the buffers should not fill up in practice. We thus expect the average latency of the current system to be between 10–20 μs.

Table 4: Round-Trip Time for a 1000 B packet: overall and per system component.

Buffer Fill Level	Round-Trip Time	Core	Ring	VEH
Empty	5.5 μs	3.6 μs	1.4 μs	0.5 μs
Half	28.7 μs	9.8 μs	17.4 μs	1.5 μs
Full	51.9 μs	16 μs	33.4 μs	2.5 μs

6.4. Partial Reconfiguration Results. Table 5 lists the local reconfiguration time and the total time needed to update a VEH. The local reconfiguration time is measured from the beginning of the DMA transfer between node-local DDR2-SDRAM and ICAP and the end of the reconfiguration process. The remote update time is measured from the first reception of a bitstream packet request at the management interface until the DONE message sent by the node PRC has been received at the Master. This time includes all data transfers of bitstream data from the management station to the system using the dedicated management network interface, sending bitstream messages on the ring from Master to the VEH node and the actual device configuration time. The measurements were made using a 10G Ethernet link at 80% utilization as the management interface. The table lists the theoretical optimum time (assuming raw transmission speeds) as well as the actual measured time on the prototype machine (by using tcpdump).

We also distinguish two cases for when looking at the local reconfiguration times. On a clean shutdown (labeled "w/ SD"), an outgoing VEH is allowed to fully process the packets already present in its input queue. Without a clean shutdown ("w/o SD"), the enqueued packets are discarded when the slot is reconfigured. For the clean shutdown measurement, we assume that the receive buffer of the VEH to be replaced is half full and that the VEH is able to process data at 10 Gb/s (being conservative, since all of our current VEHs can actually handle more).

The time required for cleanly shutting down the outgoing VEH is negligible. Most of the reconfiguration time is actually taken by feeding the bitstream into the ICAP, which limits the overall reconfiguration speed. Thus, a small size of the VEHs is important for fast reconfiguration (see also Section 6.5) and justifies our approach of heterogeneously sized VEH slots (we can configure the 14 smaller VEH slots much faster than the 4 + 4 + 2 larger ones).

When looking at the total reconfiguration time including transfer of the bitstreams from the management station, the use of a 10 Gb/s management link theoretically adds only about 40% of overhead to the raw device reconfiguration time. Even in practice, the actually measured time for reconfiguring the bitstream for the largest VEH slot over the network was just 2.3 ms. Thus, the MalCoBox can be very quickly adapted to changing attack behavior even if the bitstreams are not already present in the node-local DDR2-SDRAM, but have to be fetched from the remote management station.

In [3], Baecher et al. presented results indicating that in their experiment, the honeypot received an average of 3 requests per second, with a peak of 1300 requests per second. Based on this data, MalCoBox should have sufficient headroom for very dynamic attack scenarios even if we assume that the system manages a larger attack surface (e.g., a million IP addresses). This capability will be studied in the upcoming live test of the MalCoBox system at a central Internet router.

Note that for the more common use-case of infrequent VEH updates, a 1 Gb/s Ethernet link to the management station is quite sufficient.

6.5. Impact of Data Path Width. To evaluate the impact of the 128 b data path on the VEH size, we created 64 b versions of the SIP and the Web Server VEHs and compared them to the original 128 b implementation (Table 6). Data path conversion between the NetStage core and the VEHs can be easily performed by the wrappers at the cost of a reduced throughput for the attached VEH.

The area overhead of the 128 b version is roughly 75% for the SIP VEH and 65% for the Web Server VEH. This was to be expected, since these VEHs mostly read data from the input buffer and write data to the output buffer. The area required is thus strongly related to the bus width. Together with the data path area, the BlockRAM usage is also reduced. With 64 b operation, we can now narrow the buffers and only require three BlockRAMs per wrapper instead of five for 128 b VEHs.

Given these results, the number of parallel VEHs in the system could be increased even further by using the smaller datapath width, but only at a loss of per-VEH throughput (8 Gb/s with 64 b width and 125 MHz VEH node clocks). Assuming a heterogeneous traffic distribution across all VEHs, this would not actually lead to a slow-down, since the NetStage core would keep its 20 Gb/s-capable 128 b data path width and *distribute* the traffic across multiple of the smaller-but-slower VEHs. The bottleneck would only become apparent if all traffic was to be directed at a *single* VEH, which then would not be able to keep up with the 10 Gb/s line rate.

7. Conclusion and Next Steps

With this refinement of our MalCoBox system, we have presented a scalable architecture to build a high-speed hardware-accelerated malware collection solution that offers great flexibility through partial reconfiguration and the distribution of VEHs over multiple FPGAs. In the multidevice scenario, the total amount of VEH processing power is significantly improved in contrast to the single-chip implementation, allowing us to implement even large-scale honeynets with a single appliance. A dedicated management interface allows quick updates or replacements of single vulnerability emulation handlers by loading new partial

Table 5: Slot size distribution and reconfiguration time.

Qty.	LUT/BRAM	Bitfile size	Reconfiguration time			
			Local reconfiguration		Remote reconfiguration	
			w/o SD	w/SD	Optimum	Measured
14	1440/0	59 KB	151 μs	154 μs	218 μs	574 μs
4	2304/0	119 KB	305 μs	308 μs	432 μs	1740 μs
4	2304/2	128 KB	328 μs	332 μs	465 μs	1886 μs
2	4864/0	237 KB	607 μs	610 μs	852 μs	2269 μs

Table 6: Synthesis results for 128 b and 64 b VEHs.

VEH	LUT	Reg. Bits
SIP 128 Bit	1082	358
SIP 64 Bit	619	278
Web Server 128 Bit	1026	586
Web Server 64 Bit	663	244

bitstreams, without interrupting the operation of the rest of the system.

Enabled by the high performance of the dedicated hardware, the VEHs actually performing the malware detection and extraction can contain a wide range of functionality. They can embed complex regular expression logic as well as simple request-response patterns, while still reaching the required throughput of 10 Gb/s. Furthermore, our hardware approach is resilient against compromising attacks and significantly reduces the risk of operating honeypots in a production environment.

The presented implementation of the multi-FPGA system on the BEEcube BEE3 quad-FPGA reconfigurable computing platform demonstrated the feasibility of the approach. Operators have a great flexibility to adapt the system to their needs: A tradeoff can easily be made between individual VEH complexity and total vulnerability coverage using many different VEHs just by altering the distribution of VEH slots sizes; throughput and area can be traded off by selecting between VEH implementations with 64 b and 128 b processing widths, and the overall system size can be scaled by selecting either the single-chip or the multi-FPGA approach.

We will continue our work in this area. MalCoBox is planned to be stress-tested in a real production environment connected to the Internet (e.g., university or ISP). From this, we expect to gain valuable insights on how to improve the architecture and its parameters in the future. Furthermore, we will combine the multi-FPGA system with our recent work on self-adapting by dynamic partial reconfiguration based on the observed traffic characteristics. We expect to achieve a platform that exploits many of today's cutting-edge technologies in reconfigurable computing to enable a system presenting maximal flexibility, performance, and security to the user.

Acknowledgments

This work was supported by CASED and Xilinx, Inc.

References

[1] "Internet Security Threat Report, Volume XV," Symantec, 2010, http://www.symantec.com/.

[2] "HoneyD," http://www.honeyd.org/.

[3] P. Baecher, M. Koetter, T. Holz, M. Dornseif, and F. Freiling, "The nepenthes platform: an efficient approach to collect malware," in *Recent Advances in Intrusion Detection*,, vol. 4219 of *Lecture Notes in Computer Science*, pp. 165–184, Springer, Berlin, Germany, 2006.

[4] S. Mühlbach, M. Brunner, C. Roblee, and A. Koch, "MalCoBox: designing a 10 Gb/s malware collection honeypot using reconfigurable technology," in *Proceedings of the 20th International Conference on Field Programmable Logic and Applications (FPL '10)*, pp. 592–595, IEEE Computer Society, 2010.

[5] J. W. Lockwood, N. Naufel, J. S. Turner, and D. E. Taylor, "Reprogrammable network packet processing on the field programmable port extender (FPX)," in *Proceedings of the ACM/SIGDA 9th International Symposium on Field Programmable Gate Arrays (FPGA '01)*, pp. 87–93, ACM, 2001.

[6] S. Mühlbach and A. Koch, "A dynamically reconfigured network platform for high-speed malware collection," in *Proceedings of the International Conference on ReConFigurable Computing and FPGAs (ReConFig '10)*, pp. 79–84, IEEE Computer Society, 2010.

[7] S. Mühlbach and A. Koch, "A scalable multi-FPGA platform for complex networking applications," in *Proceedings of the IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM '11)*, pp. 81–84, IEEE Computer Society, 2011.

[8] BEEcube, Inc., "BEE3 Hardware User Manual," 2008.

[9] V. Pejović, I. Kovačević, S. Bojanić, C. Leita, J. Popović, and O. Nieto-Taladriz, "Migrating a honeypot to hardware," in *Proceedings of the International Conference on Emerging Security Information, Systems, and Technologies (SECURWARE '07)*, pp. 151–156, 2007.

[10] J. W. Lockwood, N. McKeown, G. Watson et al., "NetFPGA—an open platform for gigabit-rate network switching and routing," in *Proceedingsof the IEEE International Conference on Microelectronic Systems Education: Educating Systems Designers for the Global Economy and a Secure World (MSE '07)*, pp. 160–161, IEEE Computer Society, 2007.

[11] C. Albrecht, R. Koch, and E. Maehle, "DynaCORE: a dynamically reconfigurable coprocessor architecture for network processors," in *Proceedings of the 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pp. 101–108, IEEE Computer Society, 2006.

[12] C. Kachris and S. Vassiliadis, "Analysis of a reconfigurable network processor," in *Proceedings of the 20th International Parallel and Distributed Processing Symposium (IPDPS '06)*, p. 187, IEEE Computer Society, 2006.

[13] D. Yin, D. Unnikrishnan, Y. Liao, L. Gao, and R. Tessier, "Customizing virtual networks with partial FPGA reconfiguration," *ACM SIGCOMM—Computer Communication Review*, vol. 41, pp. 57–64, 2010.

[14] S. Bourduas and Z. Zilic, "A hybrid ring/mesh interconnectfor network-on-chip using hierarchical rings for global routing," in *Proceedings of the 1st International Symposium on Networks-on-Chip (NOCS '07)*, pp. 195–204, IEEE Computer Society, 2007.

[15] C. Thacker, "DDR2 SDRAM Controller for BEE3," Microsoft Research, 2008.

[16] K. v. d. Bok, R. Chaves, G. Kuzmanov, L. Sousa, and A. v. Genderen, "FPGA reconfigurations with run-time region delimitation," in *Proceedings of the 18th Annual Workshopon Circuits, Systems and Signal Processing (ProRISC '07)*, pp. 201–207, 2007.

[17] Y. Hori, A. Satoh, H. Sakane, and K. Toda, "Bitstream encryption and authentication using AES-GCM in dynamically reconfigurable systems," in *Proceedingsof the 3rd International Workshop on Security (IWSEC '08)*, pp. 261–278, Springer, 2008.

[18] M. Miller, "Bandwidth engine serial memory chip breaks 2 billion accesses/sec," in *Proceedings of the 23rd Hot Chips Symposium*, 2011.

[19] J. T. Pawlowski, "Hybrid memory cube: breakthrough DRAM performance with a fundamentally re-architected DRAM subsystem," in *Proceedings of the 23rd Hot Chips Symposium*, 2011.

[20] M. Thumann, "Buffer Overflow in SIP Foundry's SipXtapi," 2006, http://www.securityfocus.com/archive/1/439617.

[21] D. Litchfield, "Microsoft SQL Server 2000 Unauthenticated System Compromise," http://marc.info/?l=bugtraq&m=102760196931518&w=2.

[22] Xilinx, Partial Reconfiguration User Guide, 2010.