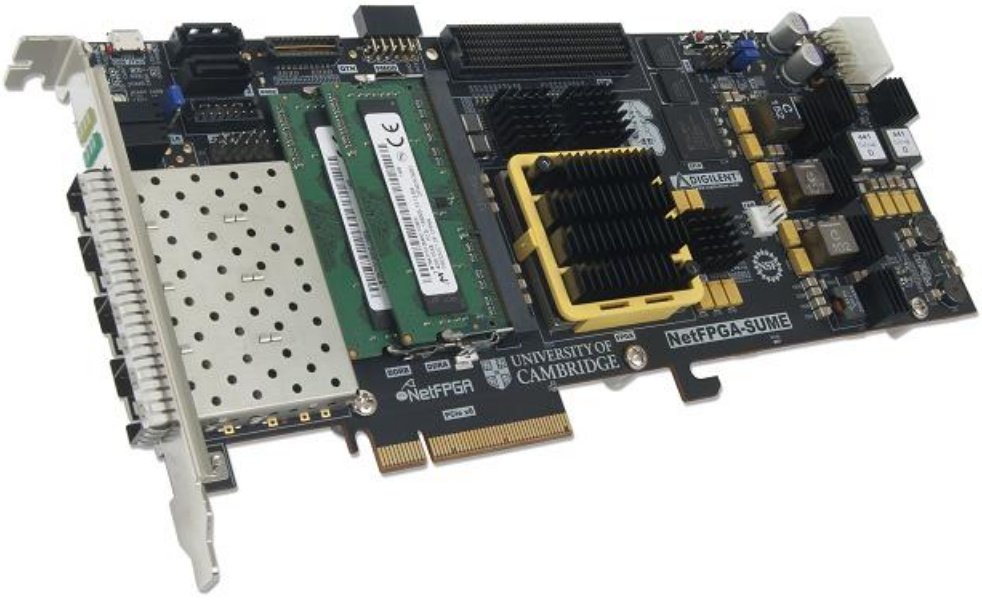
NetFPGA SUME: Toward 100 Gbps as Research Commodity

Noa Zilberman, Yury Audzevich, G. Adam Covington and Andrew W. Moore, "NetFPGA SUME: Toward 100 Gbps as Research Commodity," IEEE Micro, vol.34, no.5, pp.32-41, Sept.-Oct. 2014, doi: 10.1109/MM.2014.61



(a) 300Gb/s Ethernet Switch Using NetFPGA SUME

(b) Implementing CamCube Using NetFPGA SUME

Fig. 2. Examples of NetFPGA SUME Use Cases

also suited for implementing networking management and measurement tools, such as [1], that utilize large RAMs to implement tables for counters, lookup strings and so-on.

PCIe Host Interface: The NetFPGA SUME supports host-interface development. With the 100Gb/s physical standards still ongoing development, a host-interface capable of 100Gb/s provides the ideal prototyping vehicle for current and future interfaces. Using the uncommitted transceivers in either of the QTH and FMC expansions, permits creating two 8-lane PCIe interfaces to the host: one through the native PCIe interface and one through an expansion interface. The aggregated 128Gb/s capacity to the host (demonstrated successfully by [3]) enables exploring new and as-yet undefined physical termination standards for 100Gb/s networking.

100Gb/s Switch: In the past, the NetFPGA provided a fundamental contribution to the success of OpenFlow [14] as the initial reference platform. Switching and routing applications for 100Gb/s is a clear NetFPGA SUME application. A researcher is well placed to explore a variety of architectures in an FPGA prototyping environment. In order to construct a true non-blocking switch solution from NetFPGA SUME cards would require packet-processing at a rate of 150Mp/s for each 100Gb/s port and thus call for either a high core frequency, wide data path or combination of the two. As a result the number of physical ports available on the device is not the rate bounding element.

Using NetFPGA SUME as a true 300Gb/s fully non-blocking un-managed Ethernet switch is shown in Figure 2(a). This architecture uses a high number of high-speed serial links to deliver the required bandwidth: 100Gb/s connecting every pair of boards and providing an additional 100Gb/s port on each board. An implementation over NetFPGA SUME would use the FMC expansion interface to provide an appropriate interface: either one 100Gb/s CFP port or ten 10Gb/s SFP+ ports. The pair of 100Gb/s constituting the fabric connecting between cards can be achieved by using the transceiver resources of the PCIe connector and the QTH connector; each transceiver operating at 12.5Gb/s to achieve the per-port target bandwidth. The remaining four SFP+ ports might be used to achieve further speedup, improve signal integrity by reducing the required interface frequency, or be used to interface with the FPGA for management functions. Such a set-up might also be managed through low speed UART or I2C. This 300Gb/s switch would cost less than $5000 yet provide an extraordinary device for datacenter interconnect researchers.

A true non-blocking 300Gb/s switch requires each board to process 200Gb/s of data: 100Gb/s of inbound traffic, and 100Gb/s of outbound traffic, likely on separate datapaths. At 100Gb/s the maximal packet rate is 150Mp/s for 64B packets, however the worst case is presented by non-aligned packet sizes, e.g., 65B. Several design trade-offs exist: frequency vs. utilization vs. latency, and more. One design option may use a single data path, with 32B bus width combined with a clock rate of 450MHz. This will use less resource and will keep the latency low, yet it will pose a timing-closure challenge. An alternative design choice is to use a single data path, but as a proprietary data bus that is 96B wide and a clock rate that is only slightly more than 150MHz. This option has the

disadvantage of considerable FPGA resource utilization, but meeting timing closure would be easier. Alternatively, use multiple data paths, each 32B wide, and keep the clock frequency around 150MHz. This has a high resources utilization, and also requires additional logic for arbitration between the data paths at the output port. Using a NetFPGA SUME reference design, one can select among the options and be able to compare the performance of these three alternatives.

Physical-Layer and Media Access Control: The NetFPGA SUME permits on-FPGA reconfiguration and replacement of physical-layer and media-access controls. The expansion interfaces: FMC and QTH, each provide high-speed, standardised interfaces for researchers own daughterboard designs. Such daughter board extensions have been used to good effect for exotic interface design and are common-practice in the photonics community; permitting active and passive optical-component designs closer integration with a standard electronic interface.

Furthermore, with an ever present interest in power consumption of datacenter systems, we have treated the ability to conduct meaningful power and current analysis of a built system of high importance. NetFPGA SUME supports a purpose-specific set of power instrumentation allowing designers to study reducing of power consumption of high-speed interfaces and proving it through field measurements rather than post-synthesis analysis alone.

Interconnect: As the last example, we explore not only traditional but novel architectures with line-rate performance. Architectures that are extremely complex or require a large amount of networking equipment tend to be implemented with minimal specialist hardware. By prototyping a complete architecture, researchers can side-step limitations enforced by software-centred implementations or simulation-only studies.

In Figure 2(b) we re-create the CamCube architecture [15]. Originally six 1Gb/s links with software (host) routing; by using NetFPGA SUME could get an order of magnitude improved throughput. Figure 2(b) illustrates how N^3 NetFPGA SUME boards are connected as a $N \times N \times N$ hyper-cube: each node connects with six other nodes. NetFPGA SUME permits connecting a 40Gb/s channel to each adjacent pair of boards resulting in 240Gb/s of traffic being handled by each node.

VI. Related Work

Our approach has been to provide flexibility using an FPGA-based platform. Several such FPGA-based network-centric platforms are documented in Table I.

While the price of commercial platforms is high, ranging from $5000 to $8000, the price of a board through university affiliation programs is typically less than $2000. As the table shows, NetFPGA SUME has the most high end features. While the VC709 uses the same FPGA as the NetFPGA SUME board and same DRAM interfaces, it is a non-standard size, lacks SRAM interfaces, and has limited storage capacity. The DE5-Net board has similar DRAM access capabilities as NetFPGA SUME however, the feature set is inflexible with no additional expansion options. The NetFPGA SUME board has considerably more high-speed serial interfaces than any reference board, making it the ideal fit for high bandwidth designs.

VII. Conclusions

We present NetFPGA SUME, an FPGA-based PCIe board supporting an I/O capacity in excess of 100Gb/s provided by 30×13.1GHz transceivers, as well as SRAM and extensible DRAM memory, and a range of other useful interfaces. This is all achieved on a PCIe format board that provides a suitable HBA interface. The hardware is complemented by work done within the NetFPGA project framework providing reference software to enable researcher adoption.

NetFPGA SUME provides an important technology by serving as a platform for novel datacenter interconnect architectures, a building block for basic 100Gb/s end-host and switch research, and as a platform to explore entirely new host architectures beyond current PCIe restrictions. As a stand-alone processing unit it will enable prototype deployments otherwise too complex or too resource-intensive. As a hardware prototyping architecture, researchers are able to side-step the limitations enforced by software-centred implementations and evaluate their designs at the limits of implementation.

We have provided a brief survey of the challenges and opportunities available to researchers using hardware implementation of next generation network designs. The NetFPGA community is now set to adopt the NetFPGA SUME platform, available H2/2014, and everyone is welcome on this journey.

Acknowledgements

We thank the many people who have contributed to NetFPGA SUME project. Of particular note are the people at Xilinx: in particular, Patrick Lysaght, Michaela Blott and Cathal McCabe; the XUP programme has been a long-standing supporter of the NetFPGA and the NetFPGA SUME project is only possible with their generous support. We thank the people at Digilent Inc., in particular Clint Cole, Michael Alexander, Garrett Aufdemberg, Steven Wang and Erik Cegnar. All NetFPGA SUME pictures are courtesy of Digilent Inc. We thank Micron and Cypress Semiconductor for their generous part donations. Finally, we thank the other members of the NetFPGA project, in particular Nick McKeown at Stanford University and the entire NetFPGA team in Cambridge.

This work was jointly supported by EPSRC INTERNET Project EP/H040536/1, National Science Foundation under Grant No. CNS-0855268, and Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL), under contract FA8750-11-C-0249. The views, opinions, and/or findings contained in this report are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the National Science Foundation, Defense Advanced Research Projects Agency or the Department of Defense.

References

[1] G. Antichi et al., "OSNT: Open Source Network Tester," *IEEE Network Magazine*, September, 2014.

Platform	NetFPGA SUME	VC709	NetFPGA-10G	DE5-Net
Type	Open Source	Reference	Open Source	Reference
FPGA Type	Virtex-7	Virtex-7	Virtex-5	Stratix-V
Logical Elements	693K Logical Cells	693K Logical Cells	240K Logical Cells	622K Equivalent LEs
PCIe Hard IP	x8 Gen.3	x8 Gen.3	x8 Gen.1	x8 Gen.3
SFP+ Interfaces	4	4	4	4
Additional Serial Links	18×13.1Gb/s	10×13.1Gb/s	20×6.5Gb/s	0
Memory - On Chip	51Mb	51Mb	18Mb	50Mb
Memory - DRAM	2xDDR3 SoDIMM 4GB†, 1866MT/s	2xDDR3 SoDIMM 4GB†, 1866MT/s	4x32b RLDRAM II 576Mb, 800MT/s	2xDDR3 SoDIMM 2GB†, 1600MT/s
Memory - SRAM	27MB QDRII+, 500MHz	None	27MB QDRII, 300MHz	32MB QDRII+, 550MHz
Storage	Micro SD, 2x SATA 128MB FLASH	32MB FLASH	32MB FLASH	4x SATA 256MB FLASH
Additional Features	Expansion interface, clock recovery	Expansion interface, clock recovery		
PCI Form Factor	full-height full-length	Not compliant	full-height 3/4-length	full-height 3/4-length

TABLE I
COMPARISON BETWEEN FPGA-BASED PLATFORMS. †DENSITY PROVIDED WITH THE BOARD, EACH SUPPORTS 8GB PER SODIMM.

[2] Arista 7124FX Application Switch, *http://www.aristanetworks.com/*. [Online; accessed March 2014].
[3] Š. Friedl. et al., "Designing a Card for 100 Gb/s Network Monitoring," Tech. Rep. 7/2013, CESNET, July, 2013.
[4] J. Woodruff et al., "The CHERI Capability Model: Revisiting RISC in an Age of Risk," in *IEEE/ACM ISCA*, June, 2014.
[5] J. Wawrzynek et al., "RAMP: Research Accelerator for Multiple Processors," *IEEE Micro*, vol. 27, pp. 46–57, March, 2007.
[6] C. P. Thacker, "Improving the Future by Examining the Past: ACM Turing Award Lecture," in *IEEE/ACM ISCA*, June, 2010.
[7] I. Pratt et al., "Arsenic: a user-accessible gigabit ethernet interface," in *IEEE INFOCOM*, April, 2001.
[8] I. Leslie et al., "Fairisle: An ATM network for the local area," in *Proceedings of ACM SIGCOMM*, August, 1991.
[9] K. S. Lee et al., "SoNIC: Precise Realtime Software Access and Control of Wired Networks," in *NSDI*, pp. 213–225, April, 2013.
[10] Y. Audzevich et al., "Efficient Photonic Coding: A Considered Revision," in *ACM SIGCOMM GreenNet workshop*, pp. 13–18, August, 2011.
[11] J. W. Lockwood et al., "NetFPGA – An Open Platform for Gigabit-Rate Network Switching and Routing," *IEEE MSE*, June, 2007.
[12] M. Blott et al., "FPGA Research Design Platform Fuels Network Advances," *Xilinx Xcell Journal*, September, 2010.
[13] S. Radhakrishnan et al., "SENIC: Scalable NIC for End-Host Rate Limiting," in *USENIX NSDI*, pp. 475–488, April, 2014.
[14] N. McKeown et al., "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM CCR*, vol. 38, no. 2, pp. 69–74, 2008.
[15] H. Abu-Libdeh et al., "Symbiotic Routing in Future Data Centers," in *ACM SIGCOMM*, pp. 51–62, ACM, September, 2010.

Noa Zilberman is a Research Associate in the Systems Research Group, University of Cambridge Computer Laboratory. Since 1999 she has filled several development, architecture and managerial roles in the telecommunications and semiconductor industries. Her research interests include open-source research using the NetFPGA platform, switching architectures, high speed interfaces, Internet measurements and topology. She graduated her PhD studies in Electrical Engineering from Tel Aviv University, Israel.

Yury Audzevich is a research associate in the Computer Laboratory, Systems Research Group, University of Cambridge. His current research interests include IC design and energy-efficiency aspects in communication architectures. He has a Ph.D. in Information and Telecommunication Technologies from the University of Trento.

Adam Covington is a Research Associate in Nick McKeown's group at Stanford University. Adam has been working on the NetFPGA project since 2007. He has been helping run the NetFPGA project, both 1G and 10G, since 2009. His current research interests include reconfigurable systems, open-source hardware and software, artificial intelligence, and dynamic visualizations of large scale data. Previously, he was a Research Associate with the Reconfigurable Network Group (RNG) at Washington University in St. Louis.

Andrew W. Moore is a Senior Lecturer at the University of Cambridge Computer Laboratory in England, where he is part of the Systems Research Group working on issues of network and computer architecture. His research interests include enabling open-network research and education using the NetFPGA platform, other research pursuits include low-power energy-aware networking, and novel network and systems data-center architectures.