EPJ Web of Conferences **245**, 09004 (2020) https://doi.org/10.1051/epjconf/202024509004
CHEP 2019

EuroEXA Custom Switch: an innovative FPGA-based system for extreme scale computing in Europe

Andrea Biagioni[1], *Paolo* Cretaro[1], *Ottorino* Frezza[1], *Francesca* Lo Cicero[1], *Alessandro* Lonardo[1], *Pier Stanislao* Paolucci[1], *Luca* Pontisso[1], *Francesco* Simula[1], and *Piero* Vicini[1,*]

[1]INFN, Sezione di Roma, Italy

Abstract. EuroEXA is a major European FET research initiative that aims to deliver a proof-of-concept of a next generation Exa-scalable HPC platform. EuroEXA leverages on previous projects results (ExaNeSt, ExaNoDe and ECOSCALE) to design a medium scale but scalable, fully working HPC system prototype exploiting state-of-the-art FPGA devices that integrate compute accelerators and low-latency high-throughputnetwork.
Exascale-class systems are expected to host a very large number of computing nodes, from 10^4 up to 10^5, so that capability and performances of the interconnect architecture are critical to achieve high computing efficiency at this scale. In this perspective, EuroEXA enhances the ExaNet architecture, inherited by the ExaNeSt project, and introduces a multi-tier, hybrid topology network built on top of an FPGA-integrated Custom Switch that provides high throughput and low inter-node traffic latency for the different layers of the network hierarchy.
Deployment of a few testbeds is planned, with incremental complexity and equipped with complete software stack and runtime environment, to support the integration and test of the network design and to allow for evaluation of system performance and scalability through benchmarks based on real HPC applications. Design and integration activities are ongoing and the first small scale prototype (50 nodes) is expected to be completed in fall 2020 followed, one year later, by the deployment of the larger prototype (250/500 nodes).

1 Introduction

Nowadays, a number of technology R&D activities has been launched in Europe [1] trying to close the gap with traditional HPC providers like USA [2] and Japan [3] and more recently emerging ones like China [4].

The EU HPC strategy, funded through EuroHPC initiative, leverages on two different pillars: the first pillar is the procurement and hosting of two/three commercial pre-Exascale deployments, in order to provide the HPC community with world-level class computing systems; the second pillar aims at boosting an industry-research collaboration in order to design a new generation of Exascale systems that are to be mainly based upon European technology.

EuroEXA exploits FPGA devices, with their ensemble of either standard and custom high-performance interfaces, DSP blocks for task acceleration and a huge amount of

*e-mail: piero.vicini@roma1.infn.it

user-assigned logic cells. FPGA adoption allows us to design European innovative intellectual properties targeting either application-tailored acceleration (for high performances within the computing node) and low-latency, high-throughput custom networking (for scalability).

The EuroEXA [5] computing node, the *CRDB*, is based on a module hosting Xilinx Ultrascale+ FPGAs for application code acceleration hardware, control and network implementation, and, at a later stage, even a new project-designed, ARM-based, low power multi-core chip. The interconnect is an FPGA-based hierarchical hybrid network characterized by direct topology at blade level (16 computing nodes on a board named *Blade*) and a *Custom Switch*, implementing a mix of full-crossbar and Torus topology, for interconnection with the upper levels. EuroEXA will also introduce a new, high density liquid-cooling technology for blade systems and a new multi-rack modular assembly based on standard shipping containers in order to provide an effective solution for moving, deploying and operating large scale systems. Finally, a complete and system-optimized programming software stack is under design and a number of scientific, engineering and AI-oriented applications are used to co-design, benchmark and validate the EuroEXA hardware/software solutions.

The interconnect architecture being the focus of this document, herein we mention some solutions employed by leading competitors in HPC. For examples, the Aries interconnect in Cray XC network series [6] with its dragonfly topology [7] is regarded as one of the best alternatives. The K computer [8] at the RIKEN Advanced Institute for Computational Science (AICS) in Kobe, Japan relies on a custom interconnect — Tofu Interconnect 2 [9] arranged in a 6D-Torus topology. One the main actors in Europe is Atos, that for HPC developed the BXI [10], Bull eXascale Interconnect.

In this paper, we will introduce the EuroEXA multi-tier architecture, focusing on the design of the Custom Switch, the manager of the computing node data traffic at the lower levels of the network. A brief description of the two testbeds for the validation of the proposed architectures is in section 2 and section 3. Finally, the preliminary design of the EuroEXA Custom Switch is presented in section 4.

2 EuroEXA Testbed-1 network architecture at Mezzanine and System level

EuroEXA Testbed-1 (TB1) is the first testbed deployed. It's a small size system prototype mainly used to develop and tune the FPGA network firmware and it inherited the architecture and the computing node (the QFDB [11]) from the EU ExaNeSt project [12, 13]. In EuroEXA, we exploit the ExaNet [14] communication protocol for low-latency and high-throughput transmissions through High Speed Serial links, used to perform RDMA and shared-memory operations.

Given TB1 small number of nodes, the system do not integrate any switch device. The FPGAs within the node are "point-to-point" interconnected with an all-to-all topology. For inter-node communication of TB1 system, the QFDB provides a connector with ten bidirectional HSS links for a peak aggregated bandwidth of 100 Gbps. Four out of ten links connect neighbouring QFDBs hosted on the Mezzanine (Tier 1). The TB1 Mezzanine enables the mechanical housing of 4 QFDBs hardwired in a ring topology with two HSS links (2×10 Gbps) per edge and per direction. The remaining six links, routed through SFP+ connectors, are used to interconnect nodes residing in different Mezzanines (Tier 2).

The Torus Interface (TORUS IF in figure 1), composed by 4 HSS links, interconnects the four nodes within the same mezzanine as well as those residing on different mezzanines of the same chassis. Nine mezzanines will fit within an 11U (approximate height, the mezzanines are hosted vertically) chassis. Each chassis thus hosts 36 QFDBs — meaning 576 ARM

cores and 2.3 TB of DDR4 memory, approximately 43 cores and 210 GB of memory per 1U of cabinet height — forming a 3D-Torus, with a network diameter of 4.

A 10G Ethernet Network (10g) and an experimental Top-tier switch [15] connect nodes residing in different chassis through an interface composed by a single HSS link (10g and GEAR_IF). The main novelty of the top-tier architecture is to avoid using routing tables by devising an arithmetic routing scheme based around geographical addressing (GEAR_IF).

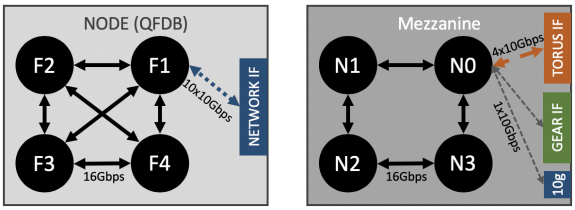

Figure 1. Testbed-1 Node and Mezzanine Topology. Four ZU9 FPGAs in an all-to-all scheme form the computing node. Four such nodes are interconnected by a ring in one mezzanine board.

3 EuroEXA Testbed-2 Network Architecture at Blade Level

Testbed-2 (TB2) is a brand new system expected to be deployed during 2020. It's based on a new computing node, the Co-design Recommended Daughter Board (CRDB), made of a Xilinx Ultrascale+ ZU9 for interconnect and compute and a Xilinx Ultrascale+ VU9 for compute acceleration. Its deployement is foreseen fro the end of 2020 Furthermore, TB2 will also introduce a new mezzanine component, the TB2 Blade, allowing the placement of 16 CRDBs interconnected using a hierarchical network with hybrid topology: all-to-all at Blade level and Torus for inter-blade connectivity.

Similar to the QFDB "Network FPGA", the CRDB integrates the same Xilinx Zynq Ultrascale+ ZU9 component to implement the computing node network peer based on the ExaNet custom protocol. Additionally, the CRDB has the same number of GTH-based serial links (10) for outer module connectivity with an expected capability of 16 Gbps on a TB2 Blade.

Unlike the TB1 multi-QFDB system, where the high number of SFP+ connectors allows for different mezzanine and chassis topologies, the mesh of CRDBs on a TB2 Blade has a fixed hierarchical topology (see figure 2): the lower network hierarchy level provides all-to-all connectivity to a group of 4 CRDBs (QUAD-DB sub system) and is implemented using 6 links per CRDB partitioned into 3 "Meshlink" channels. The remaining 4 links are bonded together in the "Uplink" channel connected to the EuroEXA Custom Switch.

The EuroEXA Custom Switch is implemented with a Xilinx Virtex VU9P FPGA. It is placed on the "Switch" board of the TB2 Blade and is connected to all uplink ports of the CRDB, providing inter-blade connectivity to higher network hierarchy levels. The Custom EuroEXA Switch is logically divided into two parts (the light blue boxes in figure 3): (i) the Blade Level EuroEXA Switch (BLES) responsible for intra-blade networking, and (ii) the Network Group EuroEXA Switch (NGES) which is responsible for routing from the BLES to other blades within the "Network Group" as well as external traffic to the Network Group Ethernet Switch.

The Network Group is a set of 8 Blades arranged into a Torus topology. NGES connectivity is mapped onto eight 100Gbps channels, made of a bonded group of four 25Gbps

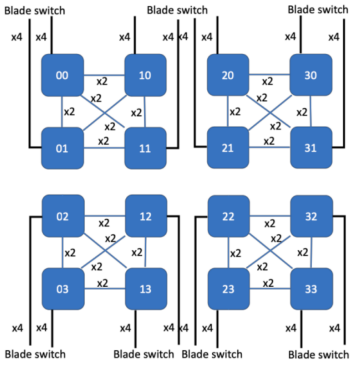

Figure 2. TB2 blade and QUAD-DB partition.

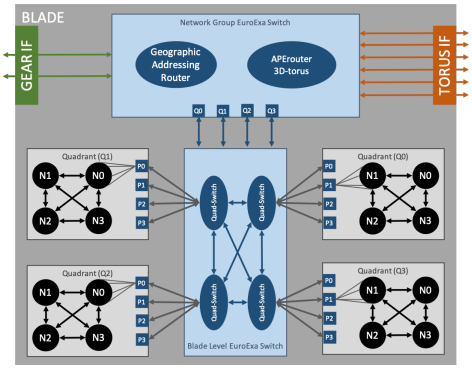

Figure 3. The EuroEXA TB2 Blade block diagram.

GTY-based links and based on QSFP28 connectors/cables mechanics. 6 out of 8 channels are used for Network Group interconnect while the remaining two channels connect the Blade to a tree of off-the-shelf Ethernet switches.

A feature and latency comparison[1] between TB1 and TB2 architectures is depicted in table 1. It is important to stress that the TB1 architecture is focused on low-latency while with the TB2 architecture EuroEXA will also step up the throughput capability. TB2 will also increase the density of computing nodes: the 36 (108) QFDBs within a chassis (rack) of TB1 will be replaced by the 128 (512) CRDBs within a Sub-Rack (Network Group) of TB2 (although the total amount of FPGAs is the same).

Table 1. The EuroEXA multi-tiered network. A comparison of TB1 and TB2 architectures

	EuroEXA Testbed-1						EuroEXA Testbed-2					
Hierarchy	**Name**	**Fanout**	**Switching**	**Topology**	**Bandwidth**	**Latency**	**Name**	**Fanout**	**Switching**	**Topology**	**Bandwidth**	**Latency**
Tier 4		>2000 Racks						250-500 Racks				
Tier 3	Rack	*x*3 Chassis	10GbE / GEAR	Fat-Tree	1x10 Gbps		Rack	*x*4 Network-Group	100GbE / GEAR	Fat-Tree	2x100 Gbps	
Tier 2	Chassis	*x*9 Mezzanines	ExaNet	3D-Torus	4x10 Gbps	400 ns per hop	Network-Group	*x*8 Blades	ExaNet	3D-Torus	4x100 Gbps	1300 ns
Tier 1	Mezzanine	*x*4 Nodes	ExaNet	Ring	1x10 Gbps	1st neigh: 400 ns	Blade	*x*16 Nodes	ExaNet	All-to-all.	2x16 Gbps	Quadrant: 300 ns
Tier 0	Node: QFDB	*x*4 FPGAs	ExaNet	All-to-All	1x16 Gbps	400 ns	Node: CRDB	*x*2 FPGAs	ExaNet		6x16 Gbps	
FPGA	Unit	ZU9						VU9+ZU9				
Core		A53						A53				

4 EuroEXA Custom Switch Preliminary design

In this section we will report on the EuroEXA FPGA-based Custom Switch activities performed during the first phase of the project. Starting from the current description of TB2 hardware and its electrical and mechanical constraints, we will discuss the rationale for VU9P FPGA selection and achievable performance expected from this component. Finally, we report on the preliminary high-level architecture of the Custom Switch and on the possible alternative design implementations.

4.1 EuroEXA Custom Switch on Xilinx Virtex Ultrascale+ FPGA

The EuroEXA Custom Switch is in charge of managing: (i) 16 ports that orchestrate communication between the Computing Nodes in the Blade (intra-Blade communication); (ii) 6 ports

[1]The TB2 latency values is preliminary and estimated using transceivers latency, measured in previous projects, applied to EuroEXA network architecture

for connection with other Blades within the Network Group (inter-blade communication); (iii) 2 ports for uplinks to Rack/System level Ethernet switch tree (inter-NG communication).

Each Computing Node (CN) provides a channel with four 16 Gbps full-duplex links, reaching 1 Tbps aggregated intra-Blade throughput ($16CN \times 4links \times 16Gbps$).

The inter-Blade aggregated throughput should be reasonably similar to the intra-Blade one in order to get a balanced network architecture. In this perspective we target the 100 Gbps performance of the 8 inter-Blade ports for 0.8 Tbps of aggregate throughput. 100 Gbps per port also allows implementing the uplink ports via the FPGA embedded Ethernet MAC and/or PHYs, a standard and cost-effective way to establish links between Blades. FPGA support for the 100G-CR4 standard allows using copper cables (a cheap option for cable) up to 5–7 meters (enough for data center rack interconnection) and QSFP28 connectors (certified up to 28 Gbps). Furthermore, by using the 100G-CR4 standard for all eight high-throughput Custom Switch channels we can avoid electrical/protocol adaptation between Ethernet Switches at Rack level and Uplink ports of the Custom Switch. As a consequence, the EuroEXA Custom Switch FPGA must provide at least 96 full duplex transceivers of which 32 ($8ports \times 4linksperport$) at 25 Gbps for network-group connectivity and the remaining 64 ($16ports \times 4linksperport$) at 16 Gbps for uplink connectivity.

Xilinx FPGAs have different classes of high speed transceivers with a max data rate of 16.3 Gbps (GTH), 32.75 Gbps (GTY) and up to 58.0 Gbps (GTM). GTY are power-efficient transceivers supporting line rates from 500 Mbps to 30.5 Gbps in UltraScale FPGAs and 32.75 Gbps in Ultrascale+ FPGAs, covering the whole range of data rates specified by the EuroEXA custom switch. Taking into account the quantity of resources per device, of embedded memory, of transceivers available per single device, the cost and the power consumption, the Virtex Ultrascale+ VU9P, with its huge amount of embedded GTY transceivers, the great deal of user available resources and large embedded memory is the best choice for the EuroEXA Custom switch implementation. Regarding the VU9P FPGA package, we selected the FLGC2104 package as the one offering just the right amount (104) of GTY transceivers.

4.2 EuroEXA Custom Switch preliminary architecture

The EuroEXA Custom Switch manages the data flow among Computing Nodes (in the following CN, aka CRDB), and its function-oriented block diagram (figure 4) can be split into I/O interfaces and switch blocks.

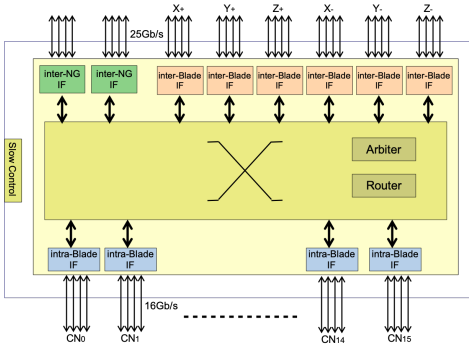

Figure 4. High-level block diagram of EuroEXA Custom Switch.

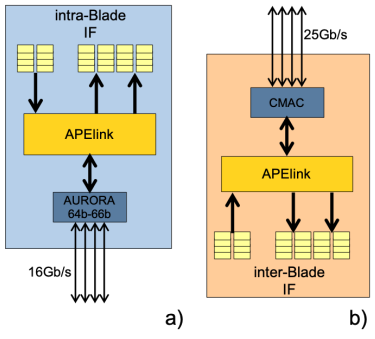

Figure 5. (a) Aurora 64B/66B-based port; (b) 100 Gigabit Ethernet port)

The I/O interfaces manage transfer of packets with format and characteristics dependent on the target devices: (i) Computing Node within the same Blade, ExaPacket [16] format - 16 intra-Blade interconnects; (ii) Remote Computing Node located in the "Network Group", ExaPacket Format - 6 inter-Blade interconnects; (iii) Network Group Ethernet Switch, Ethernet format - 2 Ethernet Switch interconnects (intra- and inter-Rack). In this paper we focus on the design of the intra/inter-Blade logic while the inter-NG module is under design by University of Manchester EuroEXA team.

The Ultrascale+ Transceiver can be configured according to industry-standard protocols, with possibly small customization for the specific system, or "from scratch" if fully custom selections are desired. For the intra-Blade channel we use the Xilinx Aurora 64B/66B configuration (figure 4). Aurora 64B/66B is a lightweight serial communication protocol for multi-Gbit links supporting 16 Gbps on GTY transceivers, with simplex or Full-duplex connections. The protocol uses 64B/66B encoding, offering improved performance because of its low (3%) transmission overhead, compared to 25% overhead for 8B/10B encoding. Aurora 64B/66B throughput is 16.375 Gbps per lane and 65.5 Gbps for a Quad (4 bonded lanes).

The Aurora-based implementation fits also perfectly for the inter-Blade communication. Moreover, we will evaluate the use of the Ultrascale+ Integrated 100G Ethernet block (figure 5) in order to increase the configurability of the design, to support both custom and Ethernet communication protocol. It supports CAUI-4 with a lane data-width of 320 bit operating at 322 MHz. Main features of Integrated 100G Ethernet core are: (i) Dynamic and static deskew support, (ii) 64B/66B decoding and encoding, and (iii) Link status and alignment monitoring.

Dataflow over all links is managed by APElink [17], a high-throughput, low-latency data transmission protocol for HPC interconnects, based on the word-stuffing technique. The switch block establishes dynamic links among the ports of the switch. It must handle more than one packet transmission at a time, managing conflicts between packets requesting the same port and ensuring deadlock avoidance.

Given the heterogeneity of implemented ports in terms of bandwidth and data-path and given the need of matching together blocks from at least three different levels in the hierarchy and the high bandwidths involved, in order to more clearly assess the various figures in performance (latency, max frequency, area) we put forward two candidate architectural solutions.

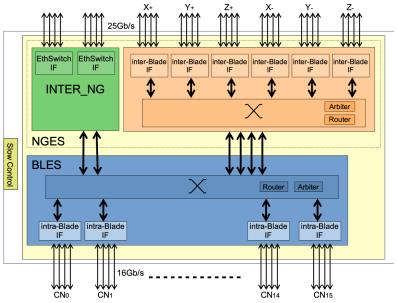

Figure 6. EuroEXA Custom Switch block diagram option 1.

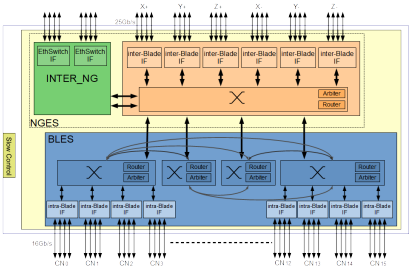

Figure 7. EuroEXA Custom Switch block diagram option 2.

.

In both architectures the BLES manages packets among local Compute Nodes while the NGES is split into 2 blocks, the first implementing the communication between Blades arranged in a 3D-torus (applying a deterministic dimension-order routing or a more sophisticated routing logic) and the latter managing the intra-Rack communication based on Ethernet switching. The first architectural option is depicted in figure 6 where the BLES block is a huge 16 + 6 ports design at different data rates (64 Gbps vs 100 Gbps) with all-to-all topology and (logic) 1-hop per transfer. The second architectural option (figure 7) sees the BLES block split into further 4 blocks, each one in charge of managing communication of Compute nodes of a single QUAD-DB sub system. These sub-blocks implement an all-to-all routing while being interconnected with a DragonFly-like topology, with a cost per port jump of 2-hops.

It is quite evident that while the first option will show lower latency for port-to-port packet transmission, the higher design complexity required to fit a 16 + 6 ports design into a single FPGA at maximum speed will negatively affect the maximum speed of the whole design. On the other hand, option 2 is slower by design (a larger number of cycles is required for port-to-port transmission) but allows for parallel arbiter/routing control phases and exhibits lower complexity due to the partition of the crossbar switch in smaller 4 sub-switch blocks. As a consequence, the comparison of the two architectural options and the selection of the most effective solution requires a further phase of exploration and evaluation of FPGA technology limits (design speed and resource use optimization) that will be performed by testing a reduced number of representative portions of the whole design.

5 Conclusion

This paper introduces the EuroEXA interconnect architecture and in particular the design of the innovative Custom Switch, the component in charge of network traffic handling at Blade and at Network Group levels. The inner components of Custom Switch, several BLES and NGES releases as well as the intra- and inter-blade channels, were successfully tested on commercial development kits to speed up the integration phase on the EuroEXA prototypes. We got very encouraging results in term of a) single (and bonded) link speed, which approaches the expected peak, and b) moderate resources occupancy. During 2020, we will finalize the design of network IPs and their porting on project designed hardware (Blade and Custom Switch PCB) in order to complete the deployment of the EuroEXA TB2 testbed. A large scale testbed, integrating the final and optimized release of the Custom Switch, is expected to be deployed in the second half of 2021.

6 Acknowledgment

This research has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 754337.

References

[1] *Eurohpc*, accessed: 08/May/2019, `https://eurohpc-ju.europa.eu/index.html`

[2] *Exascale computing project*, accessed: 08/May/2019, `https://exascaleproject.org/`

[3] *Exascale supercomputer project (riken)*, accessed: 08/May/2019, `http://www.riken.jp/en/research/rikenresearch/perspectives/2019spring/`

[4] *China's exascale supercomputer operational by 2020*, accessed: 08/May/2019, `http://english.cas.cn/newsroom/china_research/201606/t20160616_164450.shtml`

[5] *Euroexa website*, accessed: 05/Apr/2020, `https://euroexa.eu/`

[6] B. Alverson, E. Froese, L. Kaplan, D. Roweth, *Cray XC ® Series Network* (2012)

[7] J. Kim, W.J. Dally, S. Scott, D. Abts, *Technology-Driven, Highly-Scalable Dragonfly Topology*, in *2008 International Symposium on Computer Architecture* (2008), pp. 77–88, ISSN 1063-6897

[8] M. Yokokawa, F. Shoji, A. Uno, M. Kurokawa, T. Watanabe, *The K computer: Japanese next-generation supercomputer development project*, in *IEEE/ACM International Symposium on Low Power Electronics and Design* (2011), pp. 371–372, ISSN Pending

[9] Y. Ajima, T. Inoue, S. Hiramoto, S. Uno, S. Sumimoto, K. Miura, N. Shida, T. Kawashima, T. Okamoto, O. Moriyama et al., *Tofu Interconnect 2: System-on-Chip Integration of High-Performance Interconnect*, in *Supercomputing*, edited by J.M. Kunkel, T. Ludwig, H.W. Meuer (Springer International Publishing, Cham, 2014), pp. 498–507, ISBN 978-3-319-07518-1

[10] S. Derradji, T. Palfer-Sollier, J. Panziera, A. Poudes, F.W. Atos, *The BXI Interconnect Architecture*, in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects* (2015), pp. 18–25, ISSN 1550-4794

[11] F. Chaix, A. Ioannou, N. Kossifidis, N. Dimou, G. Ieronymakis, M. Marazakis, V. Papaefstathiou, V. Flouris, M. Ligerakis, G. Ailamakis et al., *Implementation and Impact of an Ultra-Compact Multi-FPGA Board for Large System Prototyping*, in *2019 IEEE/ACM International Workshop on Heterogeneous High-performance Reconfigurable Computing (H2RC)* (2019), pp. 34–41

[12] M. Katevenis, N. Chrysos, M. Marazakis, I. Mavroidis, F. Chaix, N. Kallimanis, J. Navaridas, J. Goodacre, P. Vicini, A. Biagioni et al., *The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems*, in *2016 Euromicro Conference on Digital System Design (DSD)* (2016), pp. 60–67

[13] R. Ammendola, A. Biagioni, P. Cretaro, O. Frezza, F.L. Cicero, A. Lonardo, M. Martinelli, P.S. Paolucci, E. Pastorelli, F. Simula et al., *The Next Generation of Exascale-Class Systems: The ExaNeSt Project*, in *2017 Euromicro Conference on Digital System Design (DSD)* (2017), pp. 510–515

[14] M. Katevenis, R. Ammendola, A. Biagioni, P. Cretaro, O. Frezza, F.L. Cicero, A. Lonardo, M. Martinelli, P.S. Paolucci, E. Pastorelli et al., Microprocessors and Microsystems **61**, 58 (2018)

[15] C. Concatto, J.A. Pascual, J. Navaridas, J. Lant, A. Attwood, M. Lujan, J. Goodacre, *A CAM-Free Exascalable HPC Router for Low-Energy Communications*, in *Architecture of Computing Systems – ARCS 2018*, edited by M. Berekovic, R. Buchty, H. Hamann, D. Koch, T. Pionteck (Springer International Publishing, Cham, 2018), pp. 99–111, ISBN 978-3-319-77610-1

[16] R. Ammendola, A. Biagioni, F. Capuani, P. Cretaro, G. De Bonis, F. Lo Cicero, A. Lonardo, M. Martinelli, P. Paolucci, E. Pastorelli et al., Advances in Parallel Computing **32**, 750 (2018)

[17] R. Ammendola, A. Biagioni, O. Frezza, A. Lonardo, F. Lo Cicero, P. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, P. Vicini, Journal of Instrumentation **8**, C12022 (2013)