The ExaNeSt Project: Interconnects, Storage, and Packaging for Exascale Systems

M. Katevenis[*], N. Chrysos[*], M. Marazakis[*], I. Mavroidis[*], F. Chaix[*], N. Kallimanis[*], J. Navaridas[◁], J. Goodacre[◁], P. Vicini[†], A. Biagioni[†], P. S. Paolucci[†], A. Lonardo[†], E. Pastorelli[†], F. Lo Cicero[†], R. Ammendola[†], P. Hopton[⊗], P. Coates[⊗], G. Taffoni[‡], S. Cozzini[▷], M. Kersten[∪], Y. Zhang[∪], J. Sahuquillo[±], S. Lechago[±], C. Pinto[∨], B. Lietzow[∧], D. Everett[°], and G. Perna[∩]

[*]Foundation For Research & Technology – Hellas (FORTH)
[◁]University of Manchester (UoM)
[†]National Institute for Nuclear Physics (INFN)
[⊗]Iceotope Technologies Ltd. (ICEOTOPE)
[‡]Istituto Nazionale di Astrofisica (INAF)
[▷]eXact lab srl (eXactlab)
[∪]MonetDB Solutions (MDBS)
[±]Universitat Politècnica de València (UPV)
[∨]Virtual Open Systems (VOSYS)
[∧]Fraunhofer-Gesellschaft Zur Foerderung Der Angewandten Forschung E.V (Fraunhofer)
[°]Allinea Software Ltd (Allinea)
[∩]EnginSoft S.p.A. (EnginSoft)

***Abstract*—ExaNest is one of three European projects that support a ground-breaking computing architecture for exascale-class systems built upon power-efficient 64-bit ARM processors. This group of projects share an "everything-close" and "share-anything" paradigm, which trims down the power consumption –by shortening the distance of signals for most data transfers– as well as the cost and footprint area of the installation –by reducing the number of devices needed to meet performance targets. In ExaNeSt, we will design and implement: *(i)* a physical rack prototype and its liquid-cooling subsystem providing ultra-dense compute packaging; *(ii)* a storage architecture with distributed (in-node) non-volatile memory (NVM) devices; *(iii)* a unified, low-latency interconnect, designed to efficiently uphold desired Quality-of-Service guarantees for a mix of storage with inter-processor flows; and *(iv)* efficient rack-level memory sharing, where each page is cacheable at only a single node. Our target is to test alternative storage and interconnect options on actual hardware, using real-world HPC applications. The ExaNeSt consortium brings together technology, skills, and knowledge across the entire value chain, from computing IP, packaging, and system deployment, all the way up to operating systems, storage, HPC, big data frameworks, and cutting-edge applications.**

I. Introduction

With the relentless advances in microelectronics technologies and computer architecture, the High Performance Computing (HPC) market has undergone a fundamental paradigm shift. The adoption of low-cost, Linux-based clusters extended HPC's reach from its roots in modeling and simulation of complex physical systems to a broader range of industries, from cloud computing and deep learning, to automotive and energy, many of which were originally served by datacenters.

ExaNeSt belongs to a group of recently started (Fall 2015) European projects that will support the

next step forward in this direction. Today, low-energy-consumption microprocessors (the core element of a *microserver*) dominate the embedded, smartphone and tablets markets, outnumbering x86 devices both in volume and in growth rate. If these trends continue, we can expect to see microservers benefiting from the same economies of scale that in the past favored personal computers over mainframes and, more recently, commodity clusters over custom supercomputers.

ARM is the industry leader in power-efficient processor design. ARM processors consume about 2 to 3 times less electrical energy for a given amount of computation relative to Intel-based processors, and are widely used in embedded consumer electronics, including smartphones and tablets. As a result, many research and industry programs perceive ARM-based microservers as a potential successor of x86 and POWER-based servers in hyperscale datacenters and supercomputers [1] [2] [3] and [4]. Indeed, the premonition is that using low-power processors is the only way forward towards Exascale due to its tight power budget.

Besides the power-efficiency of compute nodes, several additional challenges have to be overcome in the road towards Exascale. Modern HPC technology promises "true-fidelity" scientific simulation, enabled by the integration of huge sets of data coming from a variety of sources. As a result, the problem of *Big Data* in HPC systems is rapidly growing, fuelling a shift towards *data-centric HPC architectures*, that are expected to work on massive amounts of data, thus requiring low-latency access to fast storage. Current storage devices and interconnection networks together provide latencies on the order of hundreds of microseconds, which limit the scalability of data-hungry application models. ExaNeSt aims to address these challenges by storing data in fast storage devices, which will reside close to the processing elements.

A. In-node storage & unified interconnects

Fast non-volatile memory (NVM) (e.g. flash-based) is a key enabling technology for data-centric HPC. Aiming to avoid excessive latency and energy consumption, ExaNeSt will place these storage devices close to the compute nodes, and make them accessible through fast custom-made interconnects; for comparison, in traditional supercomputers, the storage devices are located in a central location, e.g. behind a storage area network (SAN). Placing fast storage devices close to compute elements can significantly improve the latency and the energy efficiency, as data will frequently be available in the local NVM's. Additionally, with this architecture, the capacity and I/O bandwidth of the storage subsystem scale together with compute capacity, thus securing that the system maintains its balance as we scale it out to millions of nodes.

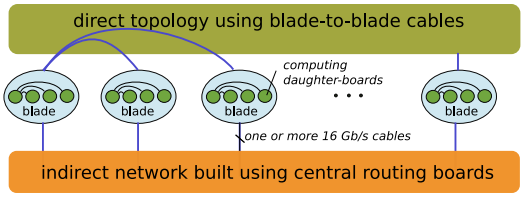

Figure 1. Networks that can be tested on Exanest's prototype: indirect topologies use central switching nodes; direct topologies have direct channels between blades (outer circles) – inner circles denote computing daughter-boards within blades; a hybrid network would have both direct and non-direct channels.

However, such a novel storage organization does not come without new challenges. To keep the system within the power and cost constraints, we envision implementing a single, *unified interconnect* that handles both storage and application traffic. Storage flows are typically bursty, causing backlogs and queuing delays inside the network;thus, they need to be dealt with carefully. A well-designed interconnect should segregate flows, through priority queues, or provide congestion control in order to protect the latency-sensitive computation messages. Additionally, the network should minimize the hops, while providing high bisection bandwidth.

Backplane interconnects can deliver high-bandwidth connectivity among the devices that reside in the same chassis or rack. In ExaNeSt, we will extend this concept, exploiting the opportunities provided by system packaging, to provide high-bandwidth connections among "neighbors" across different levels of the hierarchy (computing nodes on the same daughter-board, daughter-boards on the same blade, blades on the same chassis, etc.). Two examples of such an interconnect are shown in Fig. 1. One alternative is to have a direct topology based on the inherent networking capacity of the daughter-boards. The second alternative is to build an indirect topology based on off-the-shelf networking solutions; in ExaNeSt, we think that hybrid networks, with both direct and non-direct connections can be interesting for exascale systems.

In order to reduce the latency of (fast storage or

computation) flows, we will design user-initiated Remote Direct Memory Access (RDMA), and virtualized mailboxes, giving applications the ability to use hardware resources directly in user space. The target is to minimize the number of context switches and of data copies inside end hosts, thus enabling fast inter-process communication.

B. Rack-level shared memory

Another important feature, in order to improve the performance of many big-data and HPC applications, is the provisioning of fast, extensible DRAM memory. In ExaNeSt, the memory attached to each compute node is of modest size (tens of GBytes per compute node). In order to make large DRAM available to each compute node, we plan to enable remote memory sharing based on *Unimem*, a technology first developed within EuroServer [5], [3]. Unimem offers the ability to access areas of memory located in remote nodes. To eliminate the complexity and the costs of system-level coherence protocols [6], the Unimem architecture defines that each physical memory page can be cached at only one location. In principle, the node that caches a page can be the page owner (the node with direct access to the memory device) or any other remote node; however, in practice, it is preferred that remote nodes do not cache pages.

In ExaNeSt, we plan to extend Unimem and to show it working on a large installation with real applications. One feature that we wish to achieve is to enable a *virtual* global address space, rather than a physical one, as was the case in Euroserver. A global virtual memory page is not necessarily bound upon a specific node or a particular physical memory page. This improves security, allows page migration and can also simplify multi-programming, just as virtual memory did in the past for single node systems.

We view ExaNeSt's shared memory architecture as an alternative way to realize the objectives set out by rack disaggregation [7]: supplying enough resources to cover demand peaks without under-utilizing resources during normal conditions. Disaggregated racks break up the servers (blades) and cluster together the resulting components according to their type. In this way, we end up with blades consisting of computing devices, blades of DRAM memories, another blade holding SSD disks, etc. ExaNeSt, on the other hand, ties these units close together in nodes in order to reduce the energy consumption of data transfers, it packs many of these nodes within the rack and connects them using high-capacity interconnects.

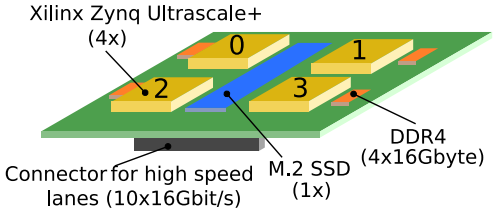

Figure 2. The ExaNeSt compute daughter-board.

Recognizing the importance of cloud computing, and its role in large HPC installations, ExaNeSt will apply *virtualization* upon 64-bit ARM-based microservers. The target is to identify and help resolve, early on, issues that may arise when deploying virtualization in actual systems.

C. ExaNeSt prototype

With all these design considerations in mind, we are presently designing two flexible prototype systems, based on Iceotope's cooling and packaging technology. The objective is to stress our complete solution (interconnect, storage, systems software) on a real implementation, using real-world HPC and big-data applications.

The basic compute unit is implemented in a daughter-board (see Fig. 2). It consists of: 4 Xilinx Ultrascale+ FPGAs, with quad ARM Cortex-A53 64-bit cores each (thus 16 cores); 16 gigabytes of DDR4 memory attached to each FPGA; and an NVM in-node SSD storage device. It will also connect to the mezzanine board of the prototype through 10 High-Speed lanes, running at up to 16 Gb/s each.

Our first prototype (Track-1) will exploit existing liquid immersion technology from Iceotope, capable to cool the thermal drive of 800W per blade. Each blade in Track 1 will host 4 to 8 compute daughter-boards (16 to 32 FPGAs). The total size of Track-1 prototype is expected to range, depending on cost, from 6 and 16 blades. The Track-2 prototype will use novel liquid cooling, developed by Iceotope during the ExaNeSt project, that will allow even denser designs, with 16 compute daughter-boards per blade.

The ExaNeSt prototype will be exercised by executing ambitious real-world applications coming from a range of scientific and industrial domains, including HPC for astrophysics and nuclear physics, neural networks for

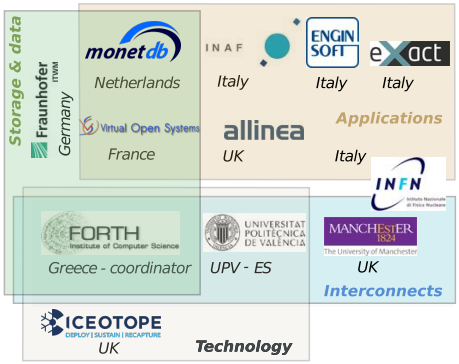

Figure 3. The ExaNeSt consortium

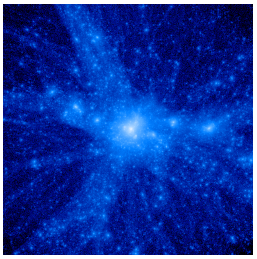

Figure 4. Results of a cosmological simulation: a galaxy cluster.

brain simulation, in-memory databases, and big-data. At the moment, these applications are being profiled in order to guide the design of the network and storage subsystems; later, a selected portion of them will be ported to the prototype, both as proof-of-concept and to evaluate our design.

D. ExaNeSt consortium and environment

The ExaNeSt consortium has been assembled in order to develop architectural solutions for HPC that will be built on an innovative technology platform. Through its six academic and six industry partners, the consortium achieves a good balance between ambitious research and pragmatic development, while covering the full R&D value chain. The members of the consortium, grouped in technical areas (which correspond to the project's work packages) are depicted in Fig. 3.

The *Foundation For Research & Technology – Hellas (FORTH)* coordinates the project; besides project management, FORTH also contributes to prototype design (technology), interconnects, and storage. In applications, key partners are *INAF* (cosmological simulations), *EnginSoft* (simulation engineering), *Allinea* (HPC code profiling and optimization), and *eXact Lab* (HPC systems). The *University of Manchester* [8], *INFN* (APEnet+ [9]), and *Univ. Polit. de Valencia* will join forces on interconnects. In storage, we have *MonetDB* (Database-as-a-Service), *Fraunhofer* (BeeGFS parallel file system), and *Virtual Open Systems* (virtualization in HPC). The two key developers of ExaNeSt prototypes are ICEOTOPE (liquid cooling) and FORTH; they will receive help on optical interconnects from UPV.

ExaNeSt is part of the broader Europe-wide ambition to pave the road to Exascale and liaises with a number of concurrent Future-and-Emerging Technologies (FET) HPC projects in order to collectively answer the HPC challenges described in the strategic vision statement of ETP4HPC [10]. Common across all projects is the technological approach for a scalable, low-energy, and economically viable solution for compute, as is being refined and realized in EuroServer [5]. Here is a summary of the projects that are aligned with this approach:

- ExaNeSt [11]: is responsible for the physical deployment characteristics to support the required compute density, along with the storage and interconnect services.
- ExaNoDe [12]: focuses on the delivery of low-energy compute elements for HPC.
- ECOSCALE [13]: focuses on integrating FPGAs and providing them as accelerators in HPC systems.

Together, these and other initiatives have the goal to deliver a European solution at the forefront of HPC.

II. Applications

The design of the ExaNeSt infrastructure will be driven and tested against scientific and industrial applications widely used in the HPC and big-data arena. The project partners have therefore selected a set of representative and ambitious test applications. Astronomers contribute with cosmological n-Body and hydrodynamical code(s) suited to perform large–scale, high–resolution numerical simulations of cosmic structures formation and evolution (see Fig. 4). In the field of Brain Simulation, a natively distributed application representative of plastic spiking neural network simulators (DP-

Figure 5. Iceotope's racks with immersion liquid-cooling.

SNN_STDP [14]) has been selected. In Engineering field, where extreme scaling would be of large benefit for scientist, engineers, and small and medium-sized enterprises (SMEs), two applications have been identified: Computational fluid dynamics (CFD) and radiation shielding. We will use one application in the area of material science simulation, one in the area of weather and climate simulation, and the MonetDB DBMS.

To benefit from the ExaNeSt infrastructure, applications must be re-designed to take advantage of the features that this provides. We expect that a new generation of exascale-ready applications will be developed during the project, and these will be used to make the final tests of the prototype.

Applications are therefore playing three important roles:

- They identify a set of initial requirements to drive the development of exascale-class platforms.
- They will be used to test and refine the ExaNeSt prototypes. Applications will also be used as benchmarks from specific domains, to provide a real comparison against competing solutions.
- Applications will be used as proof of concept to inform the design and development of systems software such as management, control, fault-tolerance, HPC communication libraries.

III. Packaging & Cooling

ExaNeSt adopts the packaging and cooling technology of our partner, Iceotope. Iceotope is leader in Totally Liquid Cooled (TLC) technology for computing infrastructures. Since the company's inception, it was recognized that liquid cooling is the future for datacenters, especially for the growth in extreme scale and density. The drivers for Iceotope's focus on TLC versus other methods of cooling include benefits in terms of efficiency, density, and total cost of ownership (TCO).

The cabinet's secondary coolant is a low cost, bespoke coolant, designed with high electrical resistance and excellent thermal properties, with over twice the heat capacity of mineral oil and half the viscosity. This coolant is circulated by fully-redundant, ultra-efficient pumps, consuming only a fraction of a percent of the energy they move. Each chassis, which is a metal enclosure connected into the rack cooling systems, can accommodate insertion of up to 9 blades, as shown in Fig. 5.

Each blade is a sealed, self-contained entity, immersed in a sophisticated, non-conductive, engineered fluid: *the primary coolant*. This coolant and the interior of the blade are designed to encourage a state of ultra-convection, known as a convective cell. This cell harnesses natural convection to rapidly moving the heat from the electronics to a secondary (rack level) coolant. When a Blade is inserted into its chassis, special valves access the cooling midplane so that the secondary coolant has access to its hotplates to draw the heat away from the inner sealed entity or chamber covering the electronics. Examples of Iceotope's primary coolant include 3M Novec and Solvay PFPE. Iceotope's secondary coolant is Iceotope Blue, formulated with a proprietary mix of minerals, anti-corrosive and anti-bacterial components.

The current Iceotope technology is designed for 72 blades per rack, at 720 Watt per blade, or 52 kW per rack, which allows for a floor density of up to 14 kW/m2. Our roadmap to exascale calls for a power density of 360 kW per rack. To achieve that within ExaNeSt, we will change the nature of the cooling cell to be "hybrid", taking advantage of both phase change and convective flow. This important innovation will require the development of some early stage technology. We will also develop a new backplane for power supply and signal I/O, and change the power distribution to 400V DC in order to be able to cope with the currents involved in such a small area.

IV. Interconnects

Current HPC systems employ one or more ultra-fast interconnects dedicated to inter-processor communication, and a separate, frequently commodity-based, network for storage traffic. The most advanced

inter-processor interconnects, although customized to provide ultra-low latencies, typically assume benign, synchronized processor traffic [15].

ExaNeSt, driven by strong power and cost incentives, focuses on a tight integration of fast storage NVMs at the node level using Unimem to improve on data locality. To fully exploit new NVMs with access latencies approaching a few tens of microseconds, we have to connect them in a low-latency, system-wide interconnect with sub-microsecond latency capabilities. In this project, we advocate the need for a unified, cross-layer optimized, low-power, hierarchical interconnect that provides equidistant communication among compute and storage devices merging inter-processor traffic with a major part of storage traffic. This consolidation of networks is expected to bring significant cost and power benefits, as the interconnect is responsible for 35% of the power budget in supercomputers and consumes power even when it idles [16].

ExaNeSt will address the different levels of the interconnect, examining suitable low-power electrical and optical technologies and appropriate topologies. A network topology matching the structure of the applications running on top of it enables maximum efficiency. In practice, interconnect topologies are severely constrained by system packaging. We address system packaging and topology selection in tandem, aiming at multitier interconnects [17], to address the disparate needs and requirements at separate building blocks inside the rack. Furthermore, we will address *inter-rack* interconnects, which span the entire system. This is considered separately because of the fundamentally disparate power and cost constraints that reign outside the enclosure of a rack. Both commodity and proprietary, electronic and optical interconnects will be examined and "judged" based on their readiness and power/performance trade offs.

The frequency and volume of checkpoint/resume traffic in exascale systems, as well as the presence of storage inside the interconnect, mandate sophisticated congestion control [18] and trouble-shooting diagnostics. Therefore, the allocation of shared resources, such as interconnect links, should be optimized for application requirements. ExaNeSt will provide quality-of-service (QoS) inside the interconnect, using hints and directions from higher layers. Small messages, such as synchronization and storage meta-data, will be prioritized appropriately. Support for QoS is required in order to isolate flows with different latency/throughput requirements and to prioritize latency-sensitive messages.

We plan to design a novel rate-based congestion control mechanism that will react to critical events, such as filled queues or links experiencing high fan-in, and will slow down or dynamically reroute the offensive flows at the sourcing host or RDMA engine. Small "synchronization" messages will be exchanged using remote load and store commands, as defined by the Unimem architecture, and in ExaNest these messages will be accelerated appropriately by the network interfaces and the interconnect. For larger messages, we will provide multi-channel RDMA engines, which can be shared among different threads, processes, VM's, or compute nodes. Our multi-channel RDMA will also provide performance isolation to its users for improved privacy and QoS.

A. Optical interconnects

Photonic interconnects are envisaged to overcome the so-called communications bottleneck of its electronic counterparts. An extensive research has been carried out regarding both on-chip and rack-to-rack photonic interconnects in terms of power, bandwidth and latency. Regarding board-to-board interconnects, it is expected that the bit rate per channel and the number of wavelength division multiplexing (WDM) channels will continue to grow in coming years, with the total capacity per link potentially reaching 1 Tb/s, using 40 channels $\times$ 25 Gb/s per channel. ExaNeSt will explore the most suitable technology in terms of efficiency and performance constraints so as to design an all-optical proof-of-concept switch. The plan is to use 2×2 and 4×4 optical switches as the main building blocks. Based on this design, we will fabricate a small-scale prototype able to fulfill demanding speed data transmission rates with low losses and low latency.

B. Resiliency

We target a unified monitoring scheme in the interconnect, which, in collaboration with appropriate agents, located at different layers of the system stack, will timely overlay critical events concerning the power consumption, the load and health of network links, and system endpoints.

Network-level tolerance and recovery from soft or hard errors is an integral part of system resiliency and a key technology in exascale systems [19], [20]. We plan

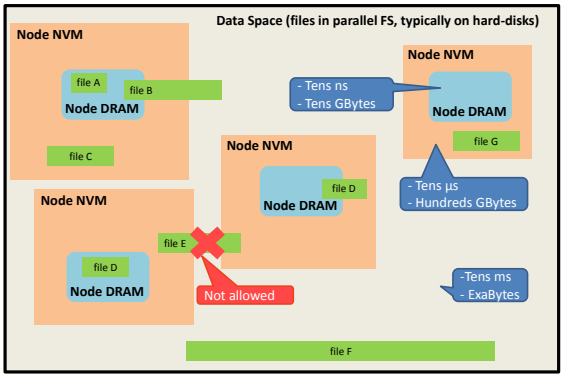

Figure 6. The ExaNeSt storage concept: DRAM main memory and NVM cache.

to leverage RDMA communication to enable dynamic routing and also to recover corrupted and undelivered packets. As the RDMA'ed packets have guaranteed space at the receiving endpoint, we can tolerate out-of-order delivery without being exposed to the danger of destination-induced deadlocks.

Of special importance is to ensure application level optimization (task placement and scheduling) in order to minimize I/O and other communication overheads by exploiting temporal and spatial locality. Finally, we plan to focus on HPC libraries (e.g. MPI collectives) and storage (e.g. metadata) traffic acceleration, using a software/hardware co-design approach. All-to-all and scatter-gather collective communications are commonly found in many HPC applications [21], [22] and become more demanding as the scale and the parallelism of the applications increase. We will study possible optimizations for HPC-relevant traffic patterns and also extend them to accelerate storage [23]. Equipping the interconnect with hardware multicast emerges as an interesting communications accelerator.

V. Storage & Data Movement

The decreasing cost of low-power fast NVM (e.g. flash-based) is changing dramatically the computing landscape. These devices promise to narrow the storage-processor performance gap with low latency (tens of microseconds vs. 10+ milliseconds), high throughput in terms of I/O operations per second (IOPS) at affordable price points, and at capacities of several hundreds GBytes versus the few GBytes of inexpensively packaged in-node DRAMs. In ExaNeSt, we envision that placing these storage devices with the compute nodes rather than in a centralized location, e.g. behind a network (SAN/NAS), is the only way forward towards Exascale. We will develop extensions to a parallel file system (BeeFS) to take advantage of such devices as cache layer. Moreover, we will devise cache maintenance protocols based on the concepts of Unimem memory consistency model. Our high-level goal is that, as long as the processors stay within the (extended) coverage offered by their local (in-node) storage devices, they achieve low-latency and low-power access to data by avoiding the movement of data across long distances. On top of that, we plan to provide replication-based resilience, protecting the filesystem by focusing on metadata integrity.

With many-thousand nodes, accessing the in-node cached data at high rates and from multiple locations makes it difficult to keep a consistent state. Furthermore, the increased possibility of failure makes things even more challenging. Figure 6 illustrates how we plan to address this challenge. File system objects are maintained in a traditional tier of hard-disk (and tape) storage devices, and in the figure they form the global persistent data space. Traditional HPC nodes would keep a small part of their data sets that can fit in their volatile memory. Whatever lies within the local boundaries of nodes is accessible at low latency and power, and does not contribute to global data movement bottlenecks. By allocating a fast storage device per node, the coverage of the nodes (the fraction of working set that can be directly accessed with the above properties), is greatly expanded. Essentially, what we will implement is a distributed storage cache, spanning across the compute nodes, that acts as a storage tier-0. At any point in time, we restrict every object to belong to no more than one cache partition, i.e., to be stored in at most one in-node NVM device. This is an extension of the Unimem model to persistent memory. When an object or a part of it resides in memory, we want to enforce that the full object can be accessed from the local storage cache. However, we still need to handle the case where the rest of the object is not accessible locally, as at most one node may keep it in its tier-0 storage. We enforce this constraint to reduce the overhead of maintaining a globally consistent cache in which objects can be present in multiple distributed storage devices.

For big-data database management systems, the presence of an NVM-based cache layer is of particular interests. The in-node NVM cache largely extends the scalability of in-memory data processing. The speed and capacity, together with its non-volatile character, make NVMs ideal intermediates between main memory and hard disks. First, during the data processing, the NVM cache layer can be used as the swap space, which is much more efficient than traditional hard disks. Second, during the downtime of a database server, the NVM layer can be used to cache the hot data set. Therefore, the NVM cache is an ideal technology to facilitate fast database restarting and migration, which in turn are important ingredients of elastic database systems. In ExaNeSt we propose heuristic NVM cache optimizations: applications will give hints of when and how certain data blocks are expected to be used –e.g., random vs. sequential accesses; temporal locality of accesses per block; read-only accesses; data blocks that should be kept together at all time; data blocks that should be pinned to certain nodes.

VI. Conclusions

ExaNeSt is currently designing one of the most advanced computing racks, consisting of densely-packed low-power 64-bit ARM processors, embedded within 16nm Xilinx FPGA SoCs, in-node distributed NVM storage, low-latency interconnects, and immersion liquid-cooling. Rack-level memory sharing, using a global virtual address space, is a key feature. The design of the ExaNeSt platform is tailored to a set of diverse and substantial applications, relevant to exascale, big data, and cloud computing.

VII. Acknowledgment

This work was carried out within the ExaNeSt project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 671553.

References

[1] R. V. Aroca and L. M. G. Gonçalves, "Towards green data centers: A comparison of x86 and ARM architectures power efficiency," *Journal of Parallel and Distributed Computing*, vol. 72, no. 12, pp. 1770–1780, 2012.

[2] R. P. Luijten and A. Doering, "The DOME embedded 64-bit microserver demonstrator," in *IC Design & Technology (ICICDT), 2013 International Conference on*. IEEE, 2013, pp. 203–206.

[3] Y. Durand, P. M. Carpenter, S. Adami, A. Bilas, D. Dutoit, A. Farcy, G. Gaydadjiev, J. Goodacre, M. Katevenis, M. Marazakis, *et al.*, "Euroserver: energy efficient node for european microservers," in *Digital System Design (DSD), 2014 17th Euromicro Conference on*. IEEE, 2014, pp. 206–213.

[4] N. Rajovic *et al.*, "Supercomputing with commodity cpus: Are mobile socs ready for hpc?" in *High Performance Computing, Networking, Storage and Analysis (SC), 2013 International Conference for*. IEEE, 2013, pp. 1–12.

[5] EuroServer, http://www.euroserver-project.eu.

[6] J. Laudon and D. Lenoski, "The SGI origin: a ccNUMA highly scalable server," in *ACM SIGARCH Computer Architecture News*, vol. 25, no. 2. ACM, 1997, pp. 241–251.

[7] S. Han *et al.*, "Network support for resource disaggregation in next-generation datacenters," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM, 2013, p. 10.

[8] J. Navaridas, M. Luján, J. Miguel-Alonso, L. A. Plana, and S. Furber, "Understanding the interconnection network of spinnaker," in *Proceedings of the 23rd international conference on Supercomputing*. ACM, 2009, pp. 286–295.

[9] R. Ammendola *et al.*, "Apenet+: a 3d torus network optimized for GPU-based hpc systems," in *Journal of Physics: Conference Series*, vol. 396, no. 4. IOP Publishing, 2012, p. 042059.

[10] ETP4HPC, http://www.etp4hpc.eu.

[11] ExaNeSt, http://www.exanest.eu.

[12] ExaNoDe, http://exanode.eu.

[13] ECOSCALE, http://www.ecoscale.eu.

[14] P. S. Paolucci *et al.*, "Dynamic many-process applications on many-tile embedded systems and HPC clusters: The EURETILE programming environment and execution platforms," *Journal of Systems Architecture*, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1383762115001423

[15] W. E. Denzel, J. Li, P. Walker, and Y. Jin, "A framework for end-to-end simulation of high-performance computing systems," *Simulation*, vol. 86, no. 5-6, pp. 331–350, 2010.

[16] T. Hoefler, "Software and hardware techniques for power-efficient hpc networking," *Computing in Science & Engineering*, vol. 12, no. 6, pp. 30–37, 2010.

[17] A. K. Kodi, B. Neel, and W. C. Brantley, "Photonic interconnects for exascale and datacenter architectures," *Micro, IEEE*, vol. 34, no. 5, pp. 18–30, 2014.

[18] N. Chrysos, L.-n. Chen, C. Kachris, and M. Katevenis, "Discharging the network from its flow control headaches: Packet drops and hol blocking," *IEEE/ACM Transactions on Networking*, 2015.

[19] M. Cuviello *et al.*, "Fault modeling and simulation for crosstalk in system-on-chip interconnects," in *Computer-Aided Design, 1999. Digest of Technical Papers. 1999 IEEE/ACM International Conference on*, Nov 1999, pp. 297–303.

[20] R. Ammendola *et al.*, "A Hierarchical Watchdog Mechanism for Systemic Fault Awareness on Distributed Systems," *Future Generation Computer Systems*, vol. 53, pp. 90 – 99, 2015.

[21] B. Arimilli *et al.*, "The percs high-performance interconnect," in *High Performance Interconnects (HOTI), 2010 IEEE 18th Annual Symposium on*, Aug 2010, pp. 75–82.

[22] M. Xie, Y. Lu, K. Wang, L. Liu, H. Cao, and X. Yang, "Tianhe-1a interconnect and message-passing services," *IEEE Micro*, vol. 32, no. 1, pp. 8–20, Jan 2012.

[23] J. Zhang, F. Ren, and C. Lin, "Modeling and understanding tcp incast in data center networks," in *INFOCOM, 2011 Proceedings IEEE*, April 2011, pp. 1377–1385.