Experiences and challenges in building next-gen optically disaggregated datacenters

(Invited Paper, Preprint version)

Andrea Reale
IBM Research - Ireland
Dublin, Republic of Ireland
realean2@ie.ibm.com

Dimitris Syrivelis
IBM Research - Ireland
Dublin, Republic of Ireland
dimitris.syrivelis@ibm.com

***Abstract*—While disaggregation has been successfully used for more than two decades to separate storage from compute (i.e., SAN/NAS systems), disaggregating tightly coupled resources like CPU, memory or accelerators has been so far considered unfeasible due to the hard dependency between their performance and the latency and bandwidth provided by their physical interconnect. We argue that modern optical networks make full datacenter disaggregation feasible in practice. We present our ongoing work on disaggregation by introducing the dReDBox architecture that completely separates datacenter resources into modular physical units interconnected via a high speed reconfigurable optical network. We discuss the current prototype and highlight challenges and work ahead.**

***Index Terms*—disaggregation, rack-scale computing, software-defined systems**

I. Introduction

In today's datacenter architectures, the server is the smallest unit of physical resource partitioning: within the boundaries of its motherboard, a server bundles a certain number of CPUs, memory and I/O peripherals. The full datacenter is then built by organizing and interconnecting servers in more or less complex topologies via high speed networks. This means that the amount and proportionality of compute and memory resources in a datacenter is directly determined by the amount and proportionality of these resources within individual servers. These are typically fixed at datacenter design time and change slowly over time with periodic hardware refresh cycles. In contrast, workloads that run in Cloud-scale datacenters present extremely dynamic characteristics in terms of the type and amount of resources they use: typical workload mixes have CPU to memory demand ratios spanning over four orders of magnitude [1], [2]. This mismatch is a serious obstacle to achieve consistent high utilization across the datacenter: in fact, as workloads are deployed, resources of different kinds tend to be depleted asymmetrically. When all the resources of a specific type in a server are completely depleted (e.g., CPUs), this makes the rest of its spare resources (e.g., memory) practically unusable, as process allocation does not span across different servers.

Resource disaggregation tackles this problem by breaking the server monolith into separate physical components, effectively creating individual pools for each type of resource. Rather than being hard-wired inside a motherboard, these resources are interconnected on-the-fly via a software-defined network to build fit-for-purpose ad-hoc computing platforms that match exactly workload requirements. Disaggregated datacenter architectures promise several advantages over traditional ones:

- Drastically improved utilization, as spare resources never become unusable.
- Reduced energy footprint, as unused resources can be independently switched off.
- Improved datacenter maintainability, as different resource types can be upgraded independently according to their own technology advancement cycles.

While disaggregation has been successfully used for more than two decades to separate storage from compute (i.e., SAN/NAS systems), disaggregating tightly coupled resources like CPU, memory or accelerators has been so far considered unfeasible due to the strict dependency between their performance and the latency and bandwidth offered by their physical interconnect. However, we argue that modern optical technologies, together with the emergence of coherent processor interfaces like OpenCAPI [3], CCIX [4], or Gen-Z [5], make complete datacenter disaggregation practically feasible today.

In this paper, we present dReDBox (disaggregated Recursive Datacenter-in-a-Box) [6]–[8], a datacenter architecture that disaggregates compute and memory resources via an high speed and software defined opto-electrical interconnect and we discuss how it has been implemented in a full-fledged hardware and software prototype (Section II). Finally, in Section III, we highlight challenges, ongoing work and future directions.

II. Disaggregating the datacenter

dReDBox disaggregates datacenter resources into composable resource units, called "bricks". Bricks can be of different types and one brick type provides a single type of resource.

This work has been partly funded by the European Commission Horizon 2020 programme, under grant agreement no. 687632.

978-1-5386-9392-6/18/$31.00 ©2018 IEEE

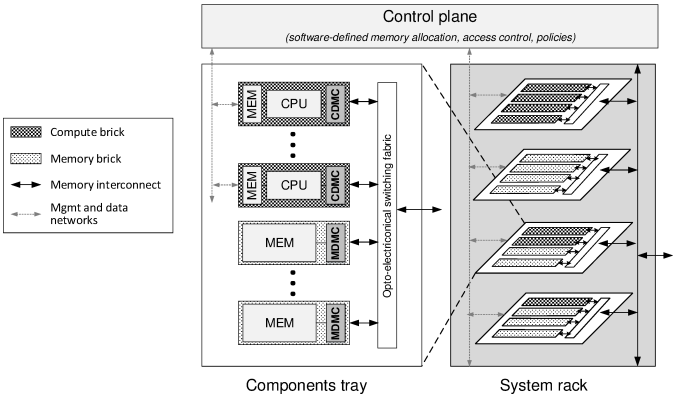

Fig. 1. dReDBox hardware/software architecture.

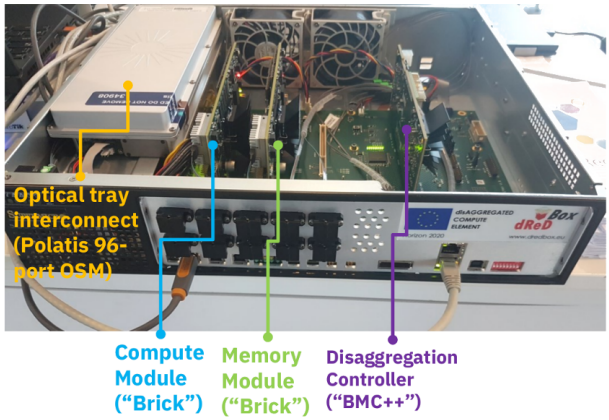

Fig. 2. dTERAY: small scale prototype of a dReDBox tray.

dReDBox defines three main types of bricks, i.e., *compute*, *memory*, and *accelerator bricks*, offering, respectively, CPUs, DRAM and FPGA accelerators to the system. For simplicity, in this paper we only focus on compute and memory bricks.

Fig. 1 shows a schematic view of the system architecture. dReDBox bricks are plugged into standard 2U rack-mountable trays; a tray can host arbitrary mixes of bricks of any type (in our system, a tray can host up to 16 resource bricks). Recursively, racks are built out of dReDBox trays, and so the full datacenter. Bricks within the same tray are interconnected via a crossbar electrical switch, while a circuit switched optical network connects bricks across different trays and racks. A software control plane oversees the system management: when a new workload arrives, the control plane configures the system software and hardware on the bricks and the network switches to create *circuits* connecting one compute brick with one or more memory bricks. This configuration process effectively creates ad-hoc physical platforms on which workloads execute. Circuit switching was chosen to favor latency over flexibility: as circuits between bricks are established we expect them to last for time intervals in the order of minutes or hours, (i.e., the full duration of their workload) so switching time is not expected to be an issue.

Unlike other attempts at memory disaggregation [9]–[11], in dReDBox, disaggregation happens at the level of Load/Store processor instructions. This means that full memory transactions are routed via the optical interconnect between their source compute brick and their destination memory brick. The important consequence of this choice is that any existing application binary can run *as-is* on a dReDBox system, with no need for any special software API or even recompilation. This is possible thanks to what we call "disaggregation memory controller" (DMC) [12], a special programmable hardware unit sitting between a brick's CPU and its transceivers (compute brick -side) and between a brick's transceivers and its memory controller (memory brick -side). Based on its configuration, the DMC intercepts transactions on the CPU bus, serializes them, and routes them towards the correct transceiver towards the destination memory brick. Configuration of the DMCs occurs at workload deployment time, and it is, again, performed by the software control plane.

On the system software side, compute bricks run a standard Linux kernel with additional modules to support DMC configuration and dynamic memory hot-plug and hot-remove [13]. Memory bricks do not need to run any OS and, in fact, do not feature a general purpose CPU.

The project consortium has implemented a full fledged hardware and software prototype of the system architecture. In the prototype, both compute and memory bricks are based on the Xilinx Zynq Ultrascale+ MPSoC (ARMv8 architecture) [14] integrated on a PCB custom designed by the project partner Sintecs [15]. Each brick has 16x 16.3 Gb/s GTH transceivers, interfaced with an embedded Luxtera LUX62608 MBO module. The tray hosting the bricks features an high-speed electrical switch-matrix for on-the-tray connectivity. For off-tray communication, connectivity is provided by a Polatis 6000 fully non-blocking all optical matrix switch module [16]. Fig. 2 shows the first small-scale prototype of this implementation hosting three bricks.

Although currently small in scale, the prototype is fully working and we are successfully using it to run real end-to-end Cloud workloads. Our goal is to collect data points that will help us understand the real potential of disaggregation beyond emulated experiments or synthetic benchmarks. Our initial exploration confirms that, while raw system memory access latency and bandwidth are obviously worse compared to those of a traditional tightly coupled memory system, this gap is significantly reduced when it comes to real end-to-end application performance. We found that this is due to three main factors: i) real applications mix I/O and computation to memory access, ii) modern large processor caches and pre-fetching units are very effective in hiding increased memory latencies and iii) similarly, burst transactions on large cache lines can significantly amortize latency towards disaggregated memory.

III. Conclusions and Ongoing Work

Resource disaggregation promises to significantly reduce the total cost of ownership of large scale datacenters by

i) improving utilization via reduced resource fragmentation, ii) reducing energy consumption thanks to increased power-off opportunities and iii) simplifying datacenter management thanks to individually manageable hardware update cycles. The dReDBox project proposes a fully-disaggregated architecture for the Cloud data-center of the future, featuring composable resource units (bricks) interconnected via a software-defined high-speed optical network. We are using a full-fledged prototype of the architecture to evaluate the potential of disaggregation, and our early results confirm that many real world applications can replace normal "local" memory with disaggregated memory with little to no performance penalty.

Our ongoing work includes increasing the prototype size to study the scalability of the system, and experimenting with novel value-add system features uniquely enabled by disaggregation, such as dynamic memory scale-up and real zero-copy virtual machine migration.

ACKNOWLEDGMENTS

dReDBox is an EU H2020 Research and Innovation Action, funded under grant agreement no. 687632. IBM Research – Ireland coordinates the project consortium which is made of other 10 academic and industrial partners, including Barcelona Supercomputing Centre, the Foundation for Research and Technology – Hellas, HUBER+SUHNER Polatis, Kinesense, Naudit, Sintecs, Telefonica, University College London, University of Thessaly, and Virtual Open Systems.

REFERENCES

[1] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *ACM Symposium on Cloud Computing (SoCC)*, San Jose, CA, USA, Oct. 2012.

[2] S. Han, N. Egi, A. Panda, S. Ratnasamy, G. Shi, and S. Shenker, "Network support for resource disaggregation in next-generation datacenters," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, ser. HotNets-XII. New York, NY, USA: ACM, 2013, pp. 10:1–10:7.

[3] OpenCAPI Consortium, "OpenCAPI Specification," Online: http://opencapi.org, 2018, accessed: August 2018.

[4] CCIX Consortium, "CCIX Specification," Online: http://wwww.ccixconsortium.com, 2018, accessed: August 2018.

[5] Gen-Z Consortium, "Gen-Z Specification," Online: http://genzconsortium.org, 2018, accessed: August 2018.

[6] K. Katrinis, D. Syrivelis, D. Pnevmatikatos, G. Zervas, D. Theodoropoulos, I. Koutsopoulos, K. Hasharoni, D. Raho, C. Pinto, F. Espina, S. Lopez-Buedo, Q. Chen, M. Nemirovsky, D. Roca, H. Klos, and T. Berends, "Rack-scale disaggregated cloud data centers: The dredbox project vision," in *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2016, pp. 690–695.

[7] G. S. Zervas, F. Jiang, Q. Chen, V. Mishra, H. Yuan, K. Katrinis, D. Syrivelis, A. Reale, D. Pnevmatikatos, M. Enrico, and N. Parsons, "Disaggregated compute, memory and network systems: A new era for optical data centre architectures," in *Optical Fiber Communication Conference*. Optical Society of America, 2017, p. W3D.4.

[8] D. Syrivelis, A. Reale, K. Katrinis, I. Syrigos, M. Bielski, D. Theodoropoulos, D. N. Pnevmatikatos, and G. Zervas, "A software-defined architecture and prototype for disaggregated memory rack scale systems," in *In Proc. of the 17th International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS) 2017*, 2017.

[9] J. Gu, Y. Lee, Y. Zhang, M. Chowdhury, and K. G. Shin, "Efficient memory disaggregation with infiniswap," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. Boston, MA: USENIX Association, 2017, pp. 649–667. [Online]. Available: https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/gu

[10] Y. Shan, S.-Y. Tsai, and Y. Zhang, "Distributed shared persistent memory," in *Proceedings of the 2017 Symposium on Cloud Computing*, ser. SoCC '17. New York, NY, USA: ACM, 2017, pp. 323–337. [Online]. Available: http://doi.acm.org/10.1145/3127479.3128610

[11] K. Lim, Y. Turner, J. R. Santos, A. AuYoung, J. Chang, P. Ranganathan, and T. F. Wenisch, "System-level implications of disaggregated memory," in *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. IEEE, 2012, pp. 1–12.

[12] D. Syrivelis, A. Reale, K. Katrinis, and C. Pinto, "A software-defined soc memory bus bridge architecture for disaggregated computing," in *In proc. of the 3rd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems (AISTECS) 2018*. New York, NY, USA: ACM, 2018.

[13] A. Reale and M. Bielski, "Memory hotplug support for arm64 — complete patchset v2," Online: https://lkml.org/lkml/2017/11/23/182, 2017, accessed: August 2018.

[14] Xilinx Inc., "Zynq UltraScale+ MPSoC," Online: https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html, 2018, accessed: September 2018.

[15] Sintecs B.V., "dReDBox: The edge of an interwoven design method." Online: https://sintecs.nl/case-dredbox/, 2017, accessed: September 2018.

[16] Huver+Suhner Polatis, "Polatis Series 6000 Optical Switch Modules," Online: http://www.polatis.com/switch-modules-for-oem-all-optical-switch-module-solutions-original-equipment-manufactures.asp, 2018, accessed: September 2018.