UCLouvain

Direct device access from the SmartNIC towards datacenter disaggregation

---

State-of-the-art

Nicolas Jeanmenne - 48741900

2025-2026

Abstract

TODO: complete abstract

This paper an is draft aiming at analyzing state-of-the-art solution for DDA from the smartNIC towards datacenter disaggregation and is not made to be submitted anywhere

State-of-the-art

Nicolas Jeanmenne

1 Introduction

For decades, servers in datacenters were designed under a monolithic model, meaning that all hardware resources are physically attached to the same frame. This model is now a bottleneck for performance, network throughput and to efficiently consume resources. In order to solve the monolithic challenge, a new architectural design has been proposed : disaggregation.

2 Definitions

2.1 Disaggregation

Disaggregation is a new model where hardware resources such as computing power, memory, accelerators, storage, ... are divided into pools (or nodes) interconnected into each other via a high-speed network. This allows resources to be allocated on demand, with better utilization and availability. There is mainly two levels of disaggregation.
Partial disaggregation separates storage components from CPU and memory which remain integrated together. This level of disaggregation has already been widely implemented since early 2020s.
Fully disaggregation regroups same type of resource into clusters which are interconnected over a network to allow communication and data exchange between them. [1]

2.2 RDMA

Remote Direct Memory Access (RDMA) is a protocol that allows to access memory from one computer into another computer without involving operating system [2]. This protocol offers high throughput and low latency, and will be referred from many articles covering this topic.

2.3 Zero-copy

Zero-copy enable the possibility to transfer data from one device to another without requiring CPU to copy the data. This avoids redundant copies, reducing CPU usage.

2.4 NVMe-oF

NVMe over-Fabric (NVMe-oF) is an extension of NVMe command set. It allows to send these commands over networked fabrics [3], [4], [5]

3 Goals

This paper will focus on fully disaggregated datacenter, especially on interaction between NVMe storage pools and SmartNIC. One of our objectives is to allow SmartNICs to communicate directly with the storage without involving any compute pool.

4 Problems and challenges

Problems can be divided into two boxes by their root cause :

- Problems due to traditional monolithic architectural server. This type of problem is the core reason of why datacenters are moving from monolithic architecture from disaggregated one.
- New challenges brought by disaggregation. Since disaggregated computing is a relatively new OS design who breaks many of assumption previously made in the research. Researchers and datacenter operators need to find creative solutions to outcome these challenges and enjoy the benefits of disaggregation.

4.1 Monolithic challenges

The main drawback with monolithic architecture (MA) is **resource stranding**. When a server exhausts one of its resource, for example memory, all other resources cannot be utilized and result in a waste of resources.
Another non-negligible issue concerning MA is failure tolerance and availability. When a piece of hardware fails this may cause the entire server to fails and become unavailable.
Finally, replacing or upgrading hardware in MA occurs maintenance that can be costly.

4.2 Disaggregation challenges

One of the main challenge in disaggregated architecture (DA) is that network bandwidth becomes a central point of communication between clusters; CPU-storage communication needs to be done in a few milliseconds while CPU-memory needs to be done in a few nanoseconds, usually 100ns [1].
Existing network stacks face significant trade-offs in terms of flexibility and performance. For example, RDMA and RDMA over Convergent Ethernet (RoCE) are high-performance protocols, but they lack flexibility because they are tightly coupled to specific hardware, making them difficult to adapt. In addition to limited flexibility, RDMA suffers from issues like head-of-line blocking, congestion, and vendor lock-in, due to hardware requirements that aren't met by commodity equipment. On the other hand, network stacks like Linux TCP offer greater flexibility but at the cost of poor performance and inefficient CPU utilization [6].

5 State-of-the-art designs and solutions

5.1 Zero-copy data path

Skiadopoulos et al. [6] designed a new prototype called *ZeroNIC* which can zero-copy data between NIC and accelerators. Its main advantage is that is it agnostic of the accelerator type and canrun on CPU, GPU, FPGA, ... NIC hardware split header and payload, transferring the latter directly to the receiver applications buffers without intermediates copies.
Concerning NVMe storage, Sun et al. [7] focused on NVMe-oF target offloading which extends zero-copy mechanism, transferring data to the DPU's host channel adapter (HCA) via peer-to-peer PCI communication (P2PDMA) and thus creating a new kind of zero-copy data path.

6 Conclusion

TODO : write conclusion

References

[1] R. Lin, Y. Cheng, M. D. Andrade, L. Wosinska, and J. Chen, "Disaggregated data centers: Challenges and trade-offs," *IEEE Communications Magazine*, vol. 58, no. 2, pp. 20–26, Feb. 2020, ISSN: 1558-1896. DOI: 10.1109/MCOM.001.1900612. Accessed: Nov. 2, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/8999422.

[2] R. J. Recio, P. R. Culley, D. Garcia, B. Metzler, and J. Hilland, "A remote direct memory access protocol specification," Internet Engineering Task Force, Request for Comments RFC 5040, Oct. 2007, Num Pages: 66. DOI: 10.17487/RFC5040. Accessed: Nov. 3, 2025. [Online]. Available: https://datatracker.ietf.org/doc/rfc5040.

[3] "NVMe over fabrics (oF) specification (historical reference only) - NVM express," Accessed: Nov. 3, 2025. [Online]. Available: https://nvmexpress.org/specification/nvme-of-specification/.

[4] "What is NVMe over fabrics? | SNIA | experts on data," Accessed: Nov. 3, 2025. [Online]. Available: https://www.snia.org/education/what-is-nvme-of.

[5] Z. Guz, H. H. Li, A. Shayesteh, and V. Balakrishnan, "NVMe-over-fabrics performance characterization and the path to low-overhead flash disaggregation," in *Proceedings of the 10th ACM International Systems and Storage Conference*, Haifa Israel: ACM, May 22, 2017, pp. 1–9, ISBN: 978-1-4503-5035-8. DOI: 10.1145/3078468.3078483. Accessed: Oct. 18, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3078468.3078483.

[6] A. Skiadopoulos et al., "High-throughput and flexible host networking for accelerated computing," presented at the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24), 2024, pp. 405–423, ISBN: 978-1-939133-40-3. Accessed: Oct. 19, 2025. [Online]. Available: https://www.usenix.org/conference/osdi24/presentation/skiadopoulos.

[7] X. Sun, M. Zhang, Y. Shan, K. Chen, J. Jiang, and Y. Wu, "Scalio: Scaling up DPU-based JBOF key-value store with NVMe-oF target offload," presented at the 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25), 2025, pp. 449–464, ISBN: 978-1-939133-47-2. Accessed: Oct. 18, 2025. [Online]. Available: https://www.usenix.org/conference/osdi25/presentation/sun.