Slashing the Disaggregation Tax in Heterogeneous Data Centers with FractOS

Lluís Vilanova
Imperial College London, London, UK

Lina Maudlej, Shai Bergman, Mark Silberstein
Technion, Haifa, Israel

Matthias Hille, Hermann Härtig
TU Dresden, Dresden, Germany

Till Miemietz, Nils Asmussen, Michael Roitzsch
Barkhausen Institut, Dresden, Germany

In pursuit of higher efficiency and lower total cost of ownership, *disaggregated heterogeneous data centers* deploy various device types, such as CPUs, accelerators, storage and memory, into separate nodes interconnected via the network, thereby facilitating specialization together with maintenance, allocation and sharing of hardware resources.

Unfortunately, applications pay a high performance tax when using current solutions because the combination of heterogeneity and disaggregation introduces redundant data transfers and control operations over the data center network. Frequent data and control transfers across multiple compute and storage devices are inherent to heterogeneous workloads. However, whereas a traditional server architecture runs such transfers over a fast dedicated local PCIe bus, a disaggregated environment instead uses a shared network with higher latency and increased performance variability (e.g., 1μs for PCIe vs. average 24μs and P99 of 40μs for RoCE RDMA in Microsoft Azure). Reducing inter-device communication overheads is thus key to attaining application performance goals in a disaggregated system.

Over the years, several system architectures have been developed to support resource disaggregation, but they are largely oblivious to the changing tradeoffs due to moving devices from a local PCIe bus to a shared network interconnect. What we need is an infrastructure to enable decentralized execution of the application logic over disaggregated heterogeneous resources, which must allow direct, peer-to-peer data and control transfer among devices and services, thereby minimizing the networking costs of disaggregation.

In this talk, we present this vision with FractOS [1], a distributed OS for disaggregated heterogeneous data centers. FractOS treats all compute and storage devices as first-class citizens like traditional CPUs, allowing them to interact directly among themselves, thus enabling fully decentralized application execution with minimal networking overheads.

Building such a system poses two key challenges. First, each device should be able to use FractOS APIs, know which task to invoke next without centralized application control, and handle exceptions during the execution. Running such a logic on each device might not be possible due to hardware constraints, such as the lack of privilege separation on GPUs, or inability to deploy user code on SSDs.

Second, allowing peer-to-peer interactions between devices without the mediation of a trusted OS is insecure. For example, references to a shared storage should not allow unauthorized data accesses. Running a security layer on each device is not viable due to the limitations above, whereas using a central security controller would not scale. FractOS offers systematic solutions to these challenges by way of an isolated OS layer, continuation-based RPCs, and a distributed capability system.

We implemented and evaluated a complete FractOS prototype and use it to build fully functional services for disaggregated GPUs and NVMe SSDs, as well as a File System service. We also implement a realistic heterogeneous face verification application that uses all of them.

References

[1] Lluís Vilanova, Lina Maudlej, Shai Bergman, Till Miemietz, Matthias Hille, Nils Asmussen, Michael Roitzsch, Hermann Härtig, Mark Silberstein. Slashing the Disaggregation Tax in Heterogeneous Data Centers with FractOS. European Conference on Computer Systems (EuroSys), pages 352–367. ACM, April 2022.