Advancing cloud with memory disaggregation

January 15 2018, by Andrea Reale

IBM Researchers – Christian Pinto, Andrea Reale, Dimitris Syrivelis, Kostas Katrinis (dReDBox Project Coordinator) and EU Programs and Contracts Manager Gal Weiss. Credit: IBM Blog Research

Here at IBM Research – Ireland, we are rethinking the very foundations on which the cloud is built. We are developing a concept and prototype for low-power and high-utilization disaggregated cloud data centres that break known boundaries, enabling the dynamic creation of fit-for-purpose computing environments from a pool of disaggregated resources.

Today's cloud data centres are implemented on a universal design axiom: the server main-board together with its hardware components form the baseline, monolithic building block that the rest of the data centre hardware/software stack design builds upon. In such high-end conventional systems, the proportionality of IT resources is fixed during design time and remains static throughout each technology refresh cycle. This comes with known ramifications in terms of low system resource utilization, costly upgrade cycles and degraded energy proportionality. The challenge in this arrangement is to be more efficient, flexible and agile with these resources.

Put simply, a data centre is a large collection of servers, each offering certain amounts of resources such as CPU cores and memory. Applications deployed to the cloud are decomposed into their basic processes, each with its own CPU and memory requirements. These processes must be started on data centre servers that have enough capacity.

CPU and memory are physically confined together within server boundaries. Because of this, an application process cannot draw resources from more than one server. The effect is fragmentation of spare CPU and memory resources. This means significant loss in economies of scale and a higher energy footprint for the cloud provider and higher service charges for the user.

Our concept is to break the physical boundaries of data centre servers by disaggregating CPUs and memory into separate physical entities. This transforms the aggregate data centre memory to a single resource pool, from which any CPU can draw resource. As requests arrive, the system can connect cores to memory on-the-fly, building platforms that match exactly their requirements.

These changes to the infrastructure are transparent to applications, so developers can keep building applications with the tools they are familiar with and deploy them to the Cloud as usual. The ultimate advantages of this disaggregation of CPU and memory are manifold. More requests can be served with the same amount of resources and memory-centric applications are no longer limited to the memory of a single server. You can also independently upgrade processors and memory, bringing agility and full modularity to the service provider.

We have built a small scale prototype of the system that proves the feasibility of our approach. So far, our experimental results have been extremely encouraging, forecasting that our architecture will help reducing up to four times the energy that today is spent on unused computing power.

We continue working to improve our technology and build larger-scale proof-of-concepts to pave the way to large scale adoption of disaggregated cloud data centers. This work spearheads the innovation in moving from today's server-as-the-unit model to a pooled-computing model; it will enable an arbitrary sizing of disaggregated IT resources that can be deployed where and when required to perfectly match cloud user requirements. This will bring the cloud to unprecedented efficiency levels while promising a drastic reduction on data centres energy footprint.

We are going to present results from this work and demonstrate our prototype at the following 2018 premier European events:

AISTECS 2018 workshop, part of the 13th HiPEAC Conference on High Performance Embedded Architectures and Compilers held in Manchester, UK, with a paper titled "A Software-defined SoC Memory Bus Bridge Architecture for Disaggregated Computing."

DATE 2018 conference and exhibition held in Dresden, Germany, with a paper and demonstration entitled "dReDBox: materializing a full-stack rack-scale system prototype of a next-generation disaggregated datacenter."

We are working under an EU funded project, dReDBox. The project is led by IBM Research Ireland, together with 10 industrial and academic partners across the European Union. As companies of all sizes continue to adopt cloud as a platform for business innovation, the impact of this type of breakthrough in Cloud's fundamental technology is far reaching.

By helping to build a more efficient Cloud, we are supporting IBM's commitment to help millions of enterprises to use the cloud to generate new business value and build better, faster and more advanced services for their clients.

More information: A Software-defined Architecture and Prototype for Disaggregated Memory Rack Scale Systems: samos-conference.com/Resources … 7/59_Final_Paper.pdf

Rack-scale disaggregated cloud data centers: The dReDBox project vision: ieeexplore.ieee.org/document/7459397/?reload=true

Provided by IBM

Citation: Advancing cloud with memory disaggregation (2018, January 15) retrieved 24 April

2024 from https://phys.org/news/2018-01-advancing-cloud-memory-disaggregation.html