Hyperconverged Infrastructure Virtualization with ReRAM

The dream of converged infrastructure is here - with hardware integration of networking, compute and storage subsystems into a single box for data centers. With this integration comes the promise of minimizing compatibility and maintenance issues, eliminating cabling and cooling to reduce costs, and reducing power consumption and floor space.

As hardware integration has taken off, vendors are now offering a more software-defined approach by integrating a software hypervisor that provides a virtualization of the three integrated functions of networking, compute and storage. This evolution is called hyper-converged infrastructure. While further delivering on the promise of integration, the hyper-converged infrastructure changes the scale-out dynamic - because the basic elements of compute and storage can't scale independently anymore. As a result, scale-out is achieved by adding a new node, which introduces new bottlenecks.

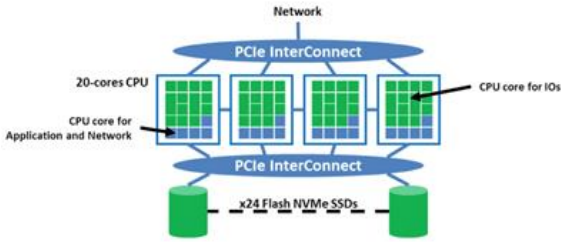
In addition to this challenge, hyperconverged applications require multi-million IO/s storage performance due to the intensive I/O workload. And yet current SSD technologies based on NAND Flash memory introduce significant latency, at 100µs to 200µs for a read I/O. To overcome the limitations of NAND Flash, IT architects developed techniques such as massive parallelization and distributed workload to compensate for those limitations by splitting storage accesses across multiple NAND Flash components. Now that the servers are moving towards hyper-convergence, it will become difficult to hide the inherent limitations of NAND Flash at the application level.

New technologies such as ReRAM are coming into the market that will slash latency to less than 10µs, resulting in new products such as NVMe SSDs. Latencies will drop even further if designers use the memory bus as the physical interface rather than PCIe. Devices that leverage the memory bus will be called NV-DIMMs, providing latencies in the microsecond range.

While the substantial performance and power benefits of ReRAM can address the storage part of the equation, this new product category will require a fresh look at the CPU/compute side as well. As system resources continue to be consumed by the storage I/O, a new architecture will be necessary to ensure compute capabilities are adequate for the application and network interface while keeping power consumption low and bandwidth high.

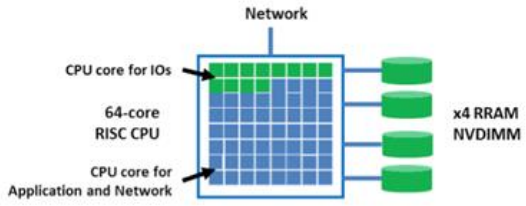
An example of this is below. The bottleneck is on the compute/storage where most of the resources are used for the storage IO/s, and then there are not enough computing capabilities for the application and for the network interface. About 3.3 cores running at full time in a high end CPU are required to manage 1 million IO/s with NVMe devices, which is an expensive power and cost budget. A typical 2U server integrates 24 SSDs, that leads to 18 million IO/s using 750 IO/s SSDs assuming the application requires a high queue depth. Therefore, 18x3.3 = 60 cores are required for the IO management, which is 75% of the resources of a high-end 4-CPU-based

architecture as illustrated in figure 1. In case the IO/s need to go over the network, the related throughput is in the range of 18 million x 4,096 x 8 = 600Gb/s, which corresponds to 15 40GbE ports.



18 million IO/s Flash-based NVMe storage system

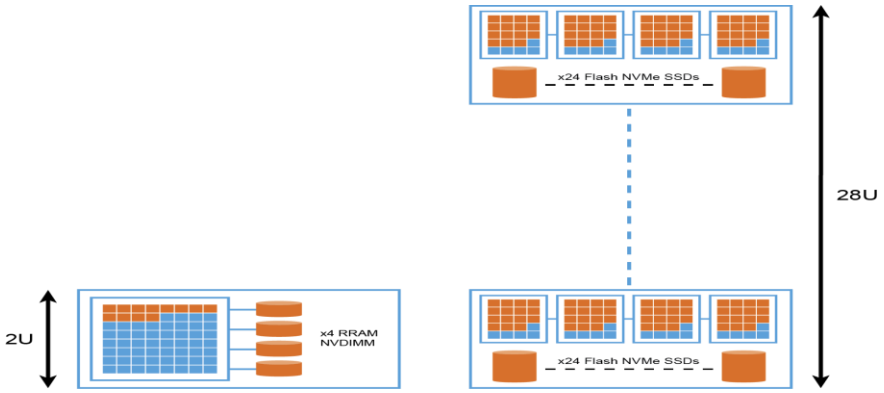
The use of RISC CPU provides enough computing capabilities for the I/O management, while maintaining low power consumption and enough bandwidth for the application and network driver. The combination of Crossbar ReRAM, through NVMe or NV-DIMM storage devices, and RISC CPUs, provides the solution in terms of IO/s and power consumption. Assuming that RISC CPU will be available at a reasonable power budget with 64 cores and 4 memory channels, we can estimate that a hyperconverged node can reach 12.5 million IO/s within a 100W power budget (figure 2). Because accessing a DIMM interface is simpler than a PCIe device, we can estimate that the storage software driver will be faster to execute compared to the NVMe driver, leading in 1 million IO/s per core for the I/O management. Due to the small form factor, about 20 of such nodes could be integrated in a 2U chassis, leading to 250 million IO/s 2U hyperconverged server, with a 2kW power budget.



12.5 million IO/s hyperconverged node

In this case, getting the IO/s over the network represents a very high bandwidth: 250 million x 4,096 x 8 = 8Tb/s, even if only 18% of the CPU resources are used for the IO management.

A virtualization use case example is a server executing 83,000 VMs in parallel (3,000 IO/s/VM). In a current Flash-based 2U hyperconverged server, integrating 24 2.5" SSDs at 750,000 IO/s per SSD, only 6,000VMs can be executed, then 14 servers are needed to execute the same VM number. ReRAM provides about 15x improvement for the I/O performance density (up to125 million IO/s/U), and performance efficiency (125,000 IO/s/W) at server level.



83,000 VMs on hyperconverged servers

In other words, for the same VM number, the user will benefit from a reduced TCO due to a more integrated solution delivering the same performance with less space, less power and fewer software licenses.

ReRAM addresses these server design challenges through low-latency/low-power storage subsystems based on this new storage class memory, removing the bottleneck from the compute/storage side.

IO Size	4kB IO	512B IO
DIMM IOPS	3M	24M
µServer IOPS (4 DIMM/µserver)	12M	96M
2U Server (20µserver/2U)	250M IOPS	2G IOPS

R&D efforts are required in the compute/network side to achieve network interfaces in the few Tb/s range, and on the software side in order to reduce the storage driver execution time. Crossbar ReRAM enables smaller IO (512B), which can be used in big data analytics and OLTP data base applications, leading to 1G IO/s per U in server.