Excelero

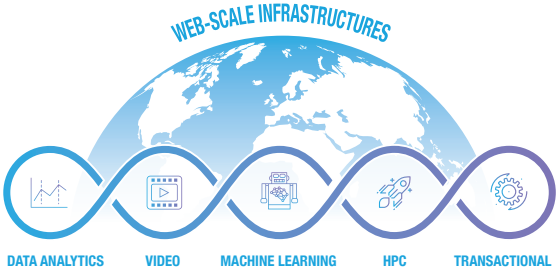
MicroTerabyte - NVMesh® - Oracle RAC Reference Architecture

TABLE OF CONTENTS

EXECUTIVE SUMMARY

CMA is a large-scale systems integrator with offices throughout the USA, specializing in delivery of large-scale database and analytics solutions for healthcare, government and financial services companies. CMA engaged with Excelero to integrate their MicroTerabyte[1] Oracle RAC solution with Excelero's NVMesh® Software-Defined Storage. The resulting solution is a highperformance, pre-installed and pre-configured, "Oracle 12c RAC cluster-in-a-box." The solution is built with very specific design goals to maximize customers' infrastructure ROI:

- Enhance the database user experience - by enabling faster and predictable response times.
- Increase multi-tenancy capabilities - by allowing multiple analysts to submit simultaneous queries
- Maximize database infrastructure flexibility - to support changing and increased workloads.
- Lower the Oracle RAC TCO - by reducing software licensing costs, optimizing hardware usage, and minimizing the datacenter footprint.



MicroTerabyte systems ship ready to deploy. The design is highly transparent, with negligible software overhead. Both the Oracle Database software and NVMesh run on Linux and the setup requires little configuration. With evolving DBaaS requirements in mind, specific effort was made to make the architecture easy to scale. In the MicroTerabyte architecture, adding RAC nodes is very straightforward and adding more NVMesh storage is equally simple. Unlike legacy designs, this takes hours, not weeks.

The MicroTerabyte example in this paper has an ultra-compact form factor. Just three RAC nodes in half a 42U rack can deliver over 60 gigabytes per second (60 GB/s) sustainable throughput. CMA compared performance results of an Excelero-based system (3-node Oracle RAC cluster) with a Dell\EMC 400K all-flash array (8-node Oracle RAC cluster) and the results were impressive: the Excelero-based solution delivered 20-22 GB/s per RAC node, while the Dell\EMC based solution only sustained 6-8 GB/s. The creation of a 10TB table space required 111 minutes on the Dell\EMC system versus 34 minutes for NVMesh. A single query scan time took 27 minutes using all 8 EMC nodes but only 3 minutes on the three-node NVMesh environment. Thus demonstrating, much more work is being accomplished with a much smaller hardware and software license footprint.

These numbers become even more impressive when you take the acquisition cost of both solutions into account: the Dell/EMC 400K solution is a multi-million dollar infrastructure, whereas the MicroTerabyte-NVMesh solution is a few hundred-thousand dollars.

[1]1 CMA MicroTerabyte information at http://cma.com/microterabyte

This paper gives an overview of the combined CMA MicroTerabyte and NVMesh solution, including a deeper dive into the design goals and architecture, a detailed overview of the benchmark results, and an overview of Excelero's NVMesh software-defined storage.

INTRODUCTION TO CMA'S MICROTERABYTE ORACLE RAC SOLUTION

CMA is a large-scale systems integrator with offices throughout the USA, specialized in delivery of large-scale database and analytics solutions for healthcare, government and financial services companies. CMA's "Products Division" builds specialty end-to-end solutions that cannot readily be found elsewhere on the market. CMA engaged with Excelero to integrate their MicroTerabyte Oracle RAC solution with Excelero's NVMesh software-defined storage. Two leading performance products combining as one solution; the result is a high-performance, pre-installed and preconfigured "Oracle 12c RAC cluster-in-a-box".

Productivity is a major challenge for Oracle RAC customers. No customer can afford two or three months to try and build an Oracle RAC solution, experiment with it, and then find out it does not work or isn't calibrated properly and won't scale. The solution CMA built with NVMesh can be delivered very quickly. Customers don't have to spend months trying to design, test, and configure the solution. With MicroTerabyte they can achieve unprecedented speed and scale in record deployment time.

The CMA MicroTerabyte solution essentially takes away the complexity of installing highperformance Oracle clusters, including the margin of error in building and developing such complex solutions. MicroTerabyte is functional and modular: customers can start small and deploy quickly. When needed, they can scale their infrastructure in a very elastic manner: expand and grow with business demand. This dramatically lowers the upfront investment.

The MicroTerabyte Oracle RAC solution with NVMesh ships, "ready to be plugged into the datacenter." CMA pre-calibrates and pretests the systems and storage for different configurations. Customers only need to focus on creating their databases and applications, then roll them out.

DBAAS AND NEXT-GENERATION DATABASE CHALLENGES

Oracle customers are looking to maximize value and productivity out of small-footprint Oracle RAC database clusters. Organizations increasingly need the ability to spin-up databases dynamically to meet changing user demands, and database performance is nowadays more critical than ever before. To meet these requirements, Oracle customers are deploying Database as a Service (DBaaS) architectures, either as on-premises infrastructure, in public cloud environments or in a hybrid setup.

Customers expect databases to be spun up very quickly, but most companies don't have budgets to build a dedicated RAC cluster for each application. Instead, they want to build a core infrastructure in which they can deploy databases as needed. This way they can fully exploit the infrastructure investment and then expand on that investment.

There is an ever-increasing demand for data: CMA maintains multiple petabytes of storage in database clusters for their customers; some individual clusters as large as hundreds of terabytes. Therefore, it is crucial to optimize the performance of these clusters: no application can sit idle to wait for a database to respond. Scale-out analytics requires results in minutes, not hours. Organizations need database platforms with the lowest latency. Elasticity, the ability to scale databases and the underlying infrastructure, is essential.

DESIGN GOALS

Clearly Software-Defined Storage implemented over SSDs is the means for achieving the performance, scale, efficiency and cost requirements of the modern data center. There are two main deployment models: converged and disaggregated.

When CMA engaged with Excelero, very specific design goals were set to maximize the infrastructure ROI. The primary objective was to minimize the software cost. Oracle licenses are expensive, so customers refrain from adding more hardware to get better performance since each additional database server CPU core requires another Oracle license. CMA wanted to shrink the cost of hardware, and with it the Oracle software licensing and, if possible, any other software costs. “Do more with less,” is the mantra, and that goal is handily achieved. Customers often run Oracle RAC clusters with four, six, or eight nodes, not because the application needs that many CPUs, but simply because of IO and bandwidth limitations per node. For example, customers who need to sustain 24 GB/s of throughput but can only deliver 4 to 6 GB/s per node would typically require 6 or 8 nodes to meet their SLAs. This is largely caused by limitations of traditional arrays and slower Fibre Channel infrastructures. However, using NVMesh, more than 20 GB/s per node can be achieved on modern server hardware. So customers can achieve higher performance with fewer nodes. This not only reduces the hardware cost and datacenter footprint, but also the Oracle RAC licensing costs.

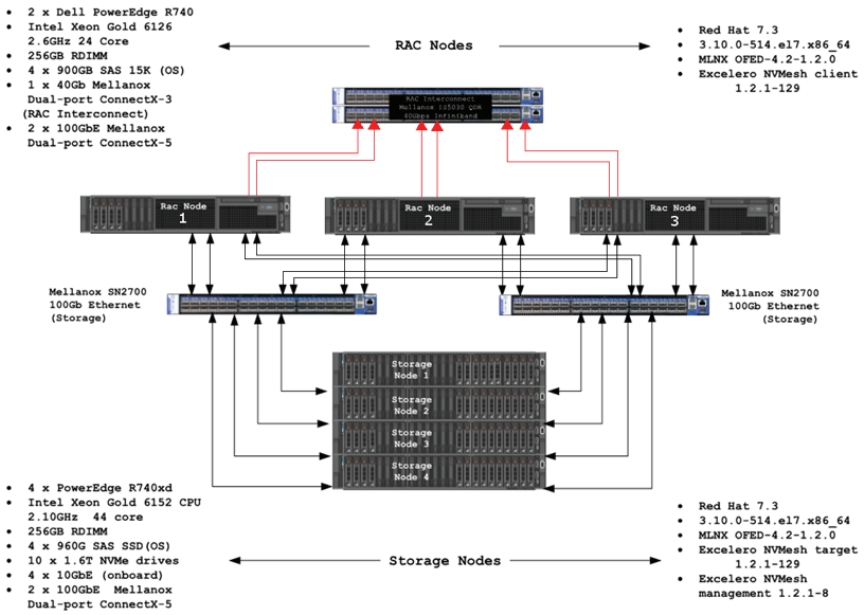
CMA also wanted to enhance the database user-experience by providing predictable, highperformance response times. Predictability is often even more important than performance: when an application delivers fast response times on some queries but slow response for others, this is considered as inconsistent performance. This hints at inherent architectural design flaws, which leverage and expect certain query plans. With Excelero’s NVMesh, CMA designed a solution that delivers extreme IOPs per node, in a simple, lightweight, and very predictable platform. Likewise, abundant IO capability allows for increased multi-tenancy with larger numbers of simultaneous queries.

Scale-out databases are under even more performance stress during maintenance operations such as backups, index builds or materialized view builds. These operations can take several hours, but applications benefit from shorter maintenance windows. Shrinking the time to create these derivative structures and complete maintenance operations is another objective for CMA. Today's database infrastructures require the flexibility to support changing and increased workloads, and the ability to create additional databases on existing platforms. MicroTerabyte allows an environment to adapt to a spike in a workload or a spike in database growth, thus ensuring productivity.

To illustrate the importance of productivity, think of a business user running a query that takes hours to complete. If that user made one small mistake (e.g. selected the wrong column) they would need to run that query again, taking several more hours. By providing a more powerful platform, that responds in minutes to seconds, customers can save valuable business time.

THE MICROTERABYTE ARCHITECTURE

A CMA MicroTerabyte base platform consists of three Oracle RAC nodes (2U Dell/Intel platform), four Dell R740xd storage nodes with NVMe drives, two 1U Mellanox IS5030 QDR switches and two Mellanox Spectrum SN2700 Ethernet switches. The IS5030 QDR switches provide the RAC interconnect, and the data IO flows through the SN2700 switches. The complete solution requires less than half of a 42U rack, but provides as much performance as 6 or 8 cabinets of gear with traditional flash and storage fabric solutions. In a nutshell, the MicroTerabyte solution delivers ultra-high performance in the smallest form factor, with minimal latency.



The RAC nodes run Red Hat Enterprise Linux v7.3 with the Mellanox OFED drivers and Excelero NVMesh initiator software. The configuration features ten 1.6TB NVMe drives in each of the four storage nodes. The storage nodes are evenly calibrated and redundantly connected into the Mellanox SN2700 switches, running 100 gigabit Ethernet. The RAC nodes use two Mellanox ConnectX-5 dual-port 100Gb/s Ethernet NICs to connect to the switches. As a result, the three RAC nodes can leverage a sustainable 60 GB/s or more; an enormous amount of bandwidth for a 20U form factor.

This starting point is just the beginning, a building block essentially: customers can expand this base infrastructure by adding RAC nodes or storage nodes. This architecture is a calibrated configuration that delivers great latency, great bandwidth, and is ready to expand. Excelero's NVMesh abstracts the NVMe capacity in the storage nodes and lets the Oracle database(s) consume the storage as if it were local storage. NVMesh comes with an intuitive software interface, making the whole infrastructure super easy to manage.

CREATING VALUE FOR CUSTOMERS

The MicroTerabyte solution delivers outstanding performance in all aspects: initial benchmarks resulted in five to ten milliseconds on very large, heavily hit platforms with large block IO. CMA further tuned the architecture until they achieved microsecond-level latencies. Latency is crucial for scale-out, high-performance databases, but so are IOPS and throughput. CMA was able to double the performance numbers of a previous architecture (that also required double the rack space). The architecture was designed to be elegant as well as simple to deploy and operate. The systems ship ready to deploy, with little software or configuration overhead. Both Oracle and NVMesh run on a standard Linux operating system, without requiring much configuration or customization. With evolving DBaaS requirements in mind, significant effort was made to enable an elastic scaleout database service. It is very straightforward to add RAC nodes or storage nodes. Adding nodes takes hours, not weeks.

MicroTerabyte has an ultra-compact form factor. Just 3 RAC nodes in half a 42U rack form factor can deliver over 60 GB/s sustainable throughput, which is more than impressive. Customers need a mere fraction of the floor space they needed previously for this level of performance. The reduced form factor, decreased power usage and hardware costs, software cost savings, combined with increased performance and you will find a massively impressive ROI. The price/performance ratio is exquisite and scaling out can be done in an economic manner as well. What is likely most important for customers is that the solution enables them to exploit cutting edge technology advancements such as NVMe flash and RDMA networking without needing any proprietary hardware. MicroTerabyte uses standard components and customers are never locked into any hardware platform: all components are interchangeable.

THE MICROTERABYTE DESIGN PROCESS

CMA initially deployed their RAC clusters on high-end storage arrays. Those systems were extremely expensive to buy, difficult to scale (forklift upgrades), required lots of datacenter space (plus power and cooling and maintenance costs) and needed many software licenses. The performance was moderate to good; nothing extraordinary, while TCO was huge.

To solve the cost vs. performance issue, CMA started to look into NVMe technology as one NVMe drive was said to deliver as many or more IOPs than an all-flash array. At first, a proprietary hardware-based NVMe solution it evaluated delivered on the performance expectations. The performance boost came at a very high cost though, including the fact that CMA and its customers were now locked into a proprietary hardware platform and the system did not scale elastically.

After the expensive hardware experience, CMA decided to build their own stack just leveraging standard hardware components, intelligent software to manage the NVMe SSDs and advanced RDMA networking. The result is the current MicroTerabyte solution with NVMesh, which offers ultra-high bandwidth and very low latency in a very tight form factor.

The NVMesh storage setup is very straightforward: customers use the GUI to create the logical volumes and attach the volumes to the RAC nodes. It operates very much like SAN storage, but without the expense and poor performance of Fibre Channel. The volumes are visualized as partitions. Storage management is much simpler than with traditional Oracle environments.

BENCHMARK RESULTS

CMA has done several benchmarks for the MicroTerabyte architecture, both off- and on-premises. In this section, we look at the results of the on-premises benchmarks. The benchmarks are truly innovative for this industry as they were designed to support massive analytics databases, with thousands of users. We compared the performance results of an Excelero-based system (3-node Oracle RAC cluster with 4 Dell\EMC R740xd) with and Dell\EMC 400K all flash array (8-node Oracle RAC cluster with 4 engine Dell\EMC VMAX). We'd like to point out again that the NVMesh-based infrastructure only has 3 RAC nodes versus 8 for the Dell\EMC-based environment. This is only possible because the Excelero software enables the application to leverage the massive NVMe throughput across the network: the Excelero-based solution delivered 20-22 GB/sec per RAC node, while the Dell\EMC based solution only sustained 6-8 GB/sec. The throughput per node is not only a crucial metric to measure performance, it is also a great one to predict how much you can save on the Oracle licensing.

Here are some key results from the benchmarks:

- The creation of a 10TB table space required 111 minutes on the Dell\EMC system versus 34 minutes for NVMesh. Note: This is a serial operation in Oracle which, if performed across all three nodes, would scale linearly. When repeated in another test, the result was even better as NVMesh enabled the creation of 30TB of table space in just over 30 minutes.
- A parallel direct path insert of 8 billion rows required 156 minutes on the Dell\EMC system versus 83 minutes for NVMesh. This test was performed on a 10TB partitioned table containing 8+ billion rows and consisted of moving the rows out of one table into another using the direct path insert in parallel mode.

- Creating a unique global partition index on the same table took only 6 minutes in the NVMesh environment versus 11 for Dell\EMC. Creating a local bitmap index 21 minutes with NVMesh versus 92 minutes for the legacy system. A single query scan time took 27 minutes using all eight Dell\EMC nodes, but only three minutes on the 3-node NVMesh environment. This is phenomenal time savings on standard database activities. Just three minutes (vs. 27) to scan a table, gives users much more flexibility in their work in terms of supporting queries and indexing. Indexing is a function of the query scan time as well.
- The maximum sustainable scan rate per RAC node was 8 GB/sec on the legacy 8-node solution, and 22.68 GB/s sustainable on the 3-node MicroTerabyte NVMesh solution. The aggregate sustainable scan rate is 30 GB/sec on the Dell\EMC 4-engine VMAX and over 60 GB/sec with NVMesh.

These numbers become even more impressive when you take the acquisition cost of both solutions into account: the Dell\EMC VMAX 400K with 4-engines and 8 RAC nodes is a multimillion dollar infrastructure, whereas the 3 RAC node MicroTerabyte-NVMesh costs only a few hundred-thousand dollars.

Benchmark	*EMC 400K (Oracle 8 node RAC cluster with 4 engine EMC VMAX)*	*Excelero (Oracle 3 node RAC cluster with 4 Dell R740xd)*
Create 10TB tablespace	111 mins	34 mins
Parallel direct path insert of 8 billion rows*	156 mins	83 mins
Create unique global partitioned index*	11 mins	6 mins
Create local bitmap index*	92 mins	21 mins
Query singletable scan time*	27 mins	3 mins
Maximum sustainable scan rate per RAC node	8 GB/Sec	22.68 GB/Sec
Aggregate sustainable scan rate	30 GB/Sec	60 GB/Sec

EXCELERO NVMESH, SOFTWARE-DEFINED STORAGE

Excelero NVMesh enables customers to design Server SAN infrastructures for the most demanding enterprise and cloud-scale applications, leveraging standard servers and multiple tiers of flash. The primary benefit of NVMesh is that it enables true converged infrastructure by logically disaggregating storage from compute.

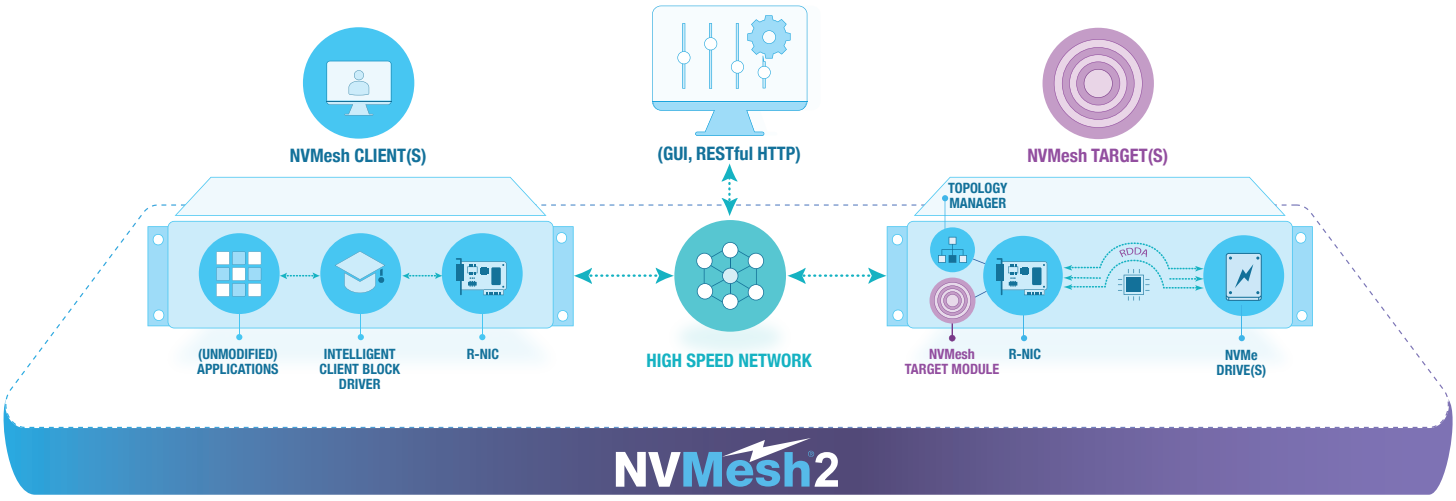
NVMesh is a software-defined block storage solution that features elastic NVMe; a distributed block layer that allows unmodified applications to utilize pooled NVMe storage devices across a network at local speeds and latencies. Distributed NVMe storage resources are pooled with the ability to create arbitrary, dynamic block volumes that can be utilized by any host running the NVMesh client. These virtual volumes can be striped, mirrored, or both, all while enjoying centralized management, monitoring and administration. In short, applications can enjoy the latency, throughput and IOPs of a local NVMe device while at the same time getting the benefits of centralized, redundant storage.

A key component of Excelero's NVMesh is the patented Remote Direct Drive Access (RDDA) functionality, which bypasses the CPU and, as a result, avoids the noisy neighbors effect common in other scale-out SDS solutions. The shift of data services from centralized CPU to complete client-side distribution enables linear

scalability, provides deterministic performance for applications and enables customers to maximize the utilization of their flash drives. NVMesh is deployed as a virtual, distributed non-volatile array and supports both converged and disaggregated architectures, giving customers full freedom in their architectural design.

HOW DOES NVMESH WORK?

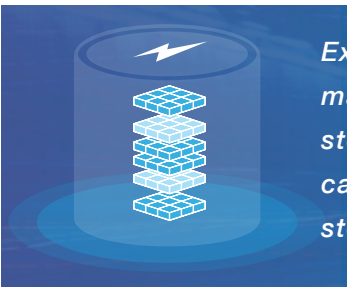
NVMesh is storage software that runs on commodity x86 Linux servers with NVMe drives. The software abstracts the NVMe drives into a single pool of high-performance, low-latency storage. Leveraging standard networking, the storage can be consumed across the network at local speeds: logical volumes of NVMe storage can be accessed across the network as if the NVMe drives were inside the server itself. NVMesh features a unique client-side architecture. In the MicroTerabyte architecture, the NVMesh client (block device initiator) is utlized on the Oracle RAC nodes. The client is anintelligent block driver that knows how to get the IO to and from the target storage nodes in the most efficient manner with the lowest possible latency. NVMesh enables true convergence on the targets, meaning it does not use any CPU on the target systems. As a result, there is no need for high-end processors, so customers can use affordable servers as storage target systems.
By pooling the NVMe and making the storage available to multiple hosts, customers can maximize utilization of their NVMe investment. Applications can use the full capacity of all NVMe drives and leverage the full performance, as if the NVMe drives were local. The architecture was designed to scale as needed, which also counts for the performance: as the cluster grows, performance scales linearly.



NVMesh has very limited hardware requirements: it runs on any modern x86 Linux server with NVMe drives and RDMA NICs (Ethernet or InfiniBand). From there, NVMesh acts as a hardware extraction layer to offer logical volumes that you define. Logical volumes can be any size, they can be striped across multiple drives, multiple hosts. Logical volumes can be created with different availability levels, with a variety of data protection schemes.

On top of the logical block volumes goes any file system, including shared and parallel file systems or, in the case of MicroTerabyte, Oracle ASM with the Oracle Clustered File System (OCFS), with multiple hosts attaching to the same block devices. The top of the stack is the application: in the case of MicroTerabyte, we are running Oracle RAC on top of ASM on top of the logical block devices, which goes all the way down to the servers.

NVMesh consists of just three main software components: the client (block initiator driver), the target module, which runs on the hosts that have the NVMe drives, and the management interface. NVMesh components can all run on one single host, or a group of hosts. An NVMesh environment can have individual clients and targets, or systems that are both clients and targets (converged). NVMesh is a complete software-defined storage stack that comes with comprehensive data protection: protection against host failures, drive failures and network failures. It was designed as a completely redundant solution.



Excelero's NVMesh is the lowest latency distributed block storage for shared NVMe on the market. It's a 100% software-defined solution that supports any hardware. Being pure block storage, NVMesh runs any local or distributed file system. NVMesh 2 adds critical sets of capabilities that make it easier for enterprises and service providers to deploy shared NVMe storage at local performance across a far wider range of network protocols and applications.

CONCLUSION

NVMesh is a 100% software solution, allowing customers to choose the hardware that best fits their requirements. CMA leveraged this flexibility to build an integrated solution that delivers extreme performance and scalability for Oracle RAC deployments while keeping the TCO as low as possible, including hardware acquisition, datacenter footprint, software licenses and management costs. Customers can scale their MicroTerabyte environments by adding nodes and drives at any time to expand their NVMe pools and logical volumes.

A huge differentiator is the fact that NVMesh uses zero CPU on the target side, which generates cost savings in many different ways: customers can use standard servers with standard CPUs – of which they need less – and they need fewer Oracle licenses to achieve better performance results.

NVMesh has a controller-less architecture, with intelligence on the client side. This takes away the controller as the traditional performance bottleneck and enables near 100% efficiency in scaling, without the need for forklift upgrades.

In a nutshell, NVMesh enables cost savings for MicroTerabyte in four ways:

- Spend less money on hardware: use standard Hardware Components, maximize NVMe efficiency
- Lower the datacenter footprint: less rack space, less power and cooling costs
- Reduce your software cost: cost-efficient SDS and fewer Oracle licenses
- Simple management: NVMesh comes with an intuitive interface, runs on standard Linux.