SUN STORAGE

An Oracle Technical White Paper
December 2011

Reference Architecture For VistaInsight for Networks on Oracle Server And Storage

Overview

As part of their ongoing technology partnership, InfoVista (IV) and Oracle ran a performance test of VistaInsight for Networks® running over Oracle Solaris and Sun hardware. The two companies shared many common objectives when starting this project:

- The most obvious was to improve the scalability and performance of the VistaInsight for Networks® (VIN) over Oracle's SPARC T-Series solution and thereby provide customers with a better price/performance ratio and a better return on investment (ROI). From the onset virtualization was considered as a promising technology to improve scalability, thereby testing VistaInsight for Networks® in the context of Oracle Solaris Containers was also a major milestone
- In addition, in response to the evolving needs of its customers, InfoVista was interested in setting new limits in terms of the workload that its application can sustain
- Finally, as the first improvements on computing-load scalability were delivered, it became obvious that the storage was the next critical component for the performance of the entire solution. A decision was then made to test the Oracle Solaris ZFS file system, the Sun ZFS Storage Appliance, and the Solid State Disk (SSD) technology from Oracle to move to the next level of performance

This document defines a reference architecture by providing detailed information about the configuration tested, the tests executed, and the results obtained. The test results clearly show that VIN takes full advantage of the server, storage and virtualization technology provided by Oracle. By leveraging the Oracle Solaris Container, Oracle Solaris ZFS, Solid State Disk and Sun ZFS Appliance storage, VIN increased its throughput performance by more than 370% - 3.7X. This solid improvement means that existing IV customers will be able to maximize the capacity of their existing hardware platforms to monitor 3.7 times more network elements without additional hardware expenditures. For new customers this performance gain translates into significant reduction in total cost of ownership (TCO) of InfoVista's solutions on Oracle servers and storage.

Application Overview

About InfoVista

InfoVista empowers telecoms operators, services providers and large enterprises to maximize business return and productivity by assuring the optimal performance of mission critical networked services, applications and infrastructures.

Eighty percent of the world’s largest service providers and more than 30 of the top 100 global corporations rely on InfoVista’s proven solutions for proactive service assurance to foresee potential service issues before they impact end users, improve business effectiveness, reduce operating risk, lower cost of operations, increase agility and create competitive advantage.

Vista Insight for Networks

Communication service providers seeking to capitalize on the growing demand for next-generation services must overcome several significant challenges that threaten their ability to deliver high quality, low-cost offerings. To attract, satisfy and retain subscribers, service providers must be able to quickly and efficiently launch new competitive offerings, such as broadband, business and mobile services, yet be able to monitor and assure the quality and availability of those services as they traverse the network infrastructure and beyond.

To address these challenges, service providers need a platform that offers them the levels of visibility required to effectively manage the complex and converged nature of today's networks and services. VistaInsight for Networks provides proactive service assurance that enables service providers and large enterprises to effectively meet and exceed performance and service quality expectations of next-generation communications technologies and services. With over 12 years of performance-focused development, VIN has demonstrated a unique scalability to align with the performance and service quality demands of a wide range of service provider products and services.

Differentiated solutions provided by VistaInsight for Networks® come mainly from out-of-the-box capability. Rich navigation features, key performance and quality indicators, reports, and integration with leading network vendors ensure a fast time-to-market of products and keep integration costs and time scales to a minimum.

InfoVista believes that proactive service assurance is not just about performance management, it brings together the concepts of performance-based troubleshooting (essential in service-oriented environments), preemptive capacity monitoring, service management and multi-tiered portal-based reporting Comprehensive use of these four concepts will enable the holistic and proactive management of end-user focused, converged IP services.

VistaInsight for Networks offers powerful self-serve capabilities that, coupled with the solution's multi-tiered reporting architecture, enable service providers to define personalized levels of access. Using informative portals, customers can investigate performance problems, make proactive decisions on capacity adjustments, verify SLAs, ensure appropriate class-of-service policies, confirm business application usage and maintain efficient use of subscribed resources. This enhances the end customer's experience and loyalty while reducing the provider's cost of customer care.

To provide holistic visibility into increasing network resources, VistaInsight for Networks takes advantage of highly scalable hardware architectures. The complexity of network performance management grows dramatically as the IP-based device population increases, requiring a hardware platform that enables VistaInsight for Networks to extend its capabilities by leveraging extra bandwidth from both computing and I/O resources.

Architecture Overview

Software Architecture

VistaInsight for Networks collects raw performance data through InfoVista Servers (IV Servers). IV Servers monitor and poll an extensive set of network devices over different protocols and interfaces. As service providers expand their network infrastructure, more devices and interfaces need to be monitored which creates the need to deploy additional IV Server instances to cope with the load.

As the raw performance data is collected, IV Servers store it in a local database that runs on the same hardware system. In production the IV Server instances can be spread across many systems to accommodate the network topology yet, with the adoption of virtualization across the industry, a growing number of customers consolidate IV Servers on a single physical system. This consolidation configuration was tested during the tests as IV Server instances were running in Oracle Solaris Containers.

Finally, when raw data has been collected and stored in an IV Server local database, it is then forwarded to a centralized server – VistaMart – that runs on a separate system. VistaMart consolidates the raw data from the different InfoVista Servers into a single database that is used for further consultation and data mining operations.

This test focused on the scalability of the IV Servers and their local databases.

Platform Architecture

To achieve scalability this test took advantage of the server and storage technologies from Oracle.

SPARC T-Series servers have the capacity to execute a very large number of simultaneous threads. This makes them natural candidates for virtualization as this technology generates a high-level of parallelism. The Sun SPARC Enterprise T5240 server was thereby chosen to host the different IV Server instances. It delivers 128 computing threads with only two UltraSPARC T2 Plus processors (i.e. 64 threads per processor).

In terms of storage I/O, the bulk of the load was generated by the IV Sever local database: ObjectStore, an object oriented transactional database. 8 IV Servers were sending collected raw data to a single ObjectStore instance.

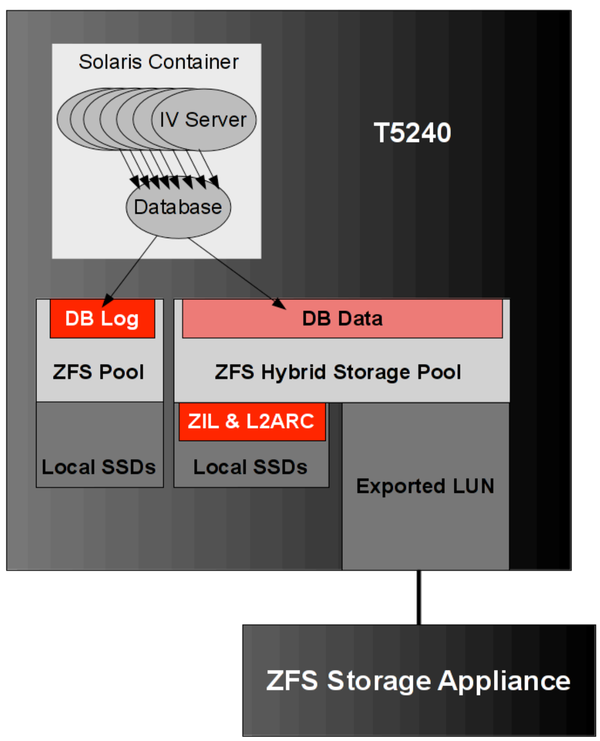
The I/O load of a transactional database produces a lot of synchronous writes. The bottleneck for the synchronous writes is focused on a small number of files—the transaction logs—that contain a small amount of data. This is a typical situation where the combination of Oracle Solaris ZFS and SSD can be used to improve performance without compromising cost. Oracle Solaris ZFS is flexible enough so that a small number of SSDs can be dedicated to—and only to—the transaction logs.

In the case of IV Server, these SSDs—provided by a Sun Storage F5100 Flash Array—were directly attached to the T5240. They behaved as local SSDs in the T5240 and were as close as possible to the ObjectStore database. The rest of the data that represented the bulk of the storage consumption, but that did not generate a high amount of critical I/O operations, was located on a price-efficient Sun ZFS Storage 7210[1] Appliance used as a SAN.

For the performance evaluation, Oracle Solaris ZFS flexibility was used in different ways: It allowed for the creation of a separate storage pool for the ObjectStore's transaction logs. Doing so segregated the critical operations—the synchronous writes—from the rest of the I/O operations: a sync write was never waiting for another operation performed on the bulk of the data to complete. In addition, since this storage pool only contained the transaction logs, its size was limited and it could fit on a small number of SSDs, hence limiting the impact on the overall cost

Oracle Solaris ZFS was also used as the file system on top of the LUN exported by the SAN. Here, the ZFS Hybrid Storage Pool architecture was leveraged to once again take advantage of the SSD performance for critical operations. When it comes to I/O operations, Oracle Solaris ZFS acts similarly to a transactional database and in addition, it comes with built-in features to easily and automatically dedicate SSDs to critical operations: these are the ZFS Intent Log (a.k.a. ZIL) and the ZFS level-3 cache for read (a.k.a. L2ARC). The storage pool dedicated to the bulk of the data ended up being a mixed of local SSDs and a remote LUN—hence the terminology of Hybrid Storage Pool.

The following diagram provides a logical view of the overall infrastructure:



Workload Description and Test Environment

Workload Description

To simulate the network devices, the following topology was used that simulated approximately 80K device instances:

- 50 Juniper ERX M320 with 1277 logical interface each
- 25 Cisco 7600 series with 100 proxies
- 10 ADVA Optical devices with 11 proxies

Monitoring this topology generated a load of 2500 samples/sec per IV Server, for a total load of 8 x 2500 = 20K samples/sec. The term samples/sec is an internal value of IV Server indicating how much data the ObjectStore server can handle in real time.

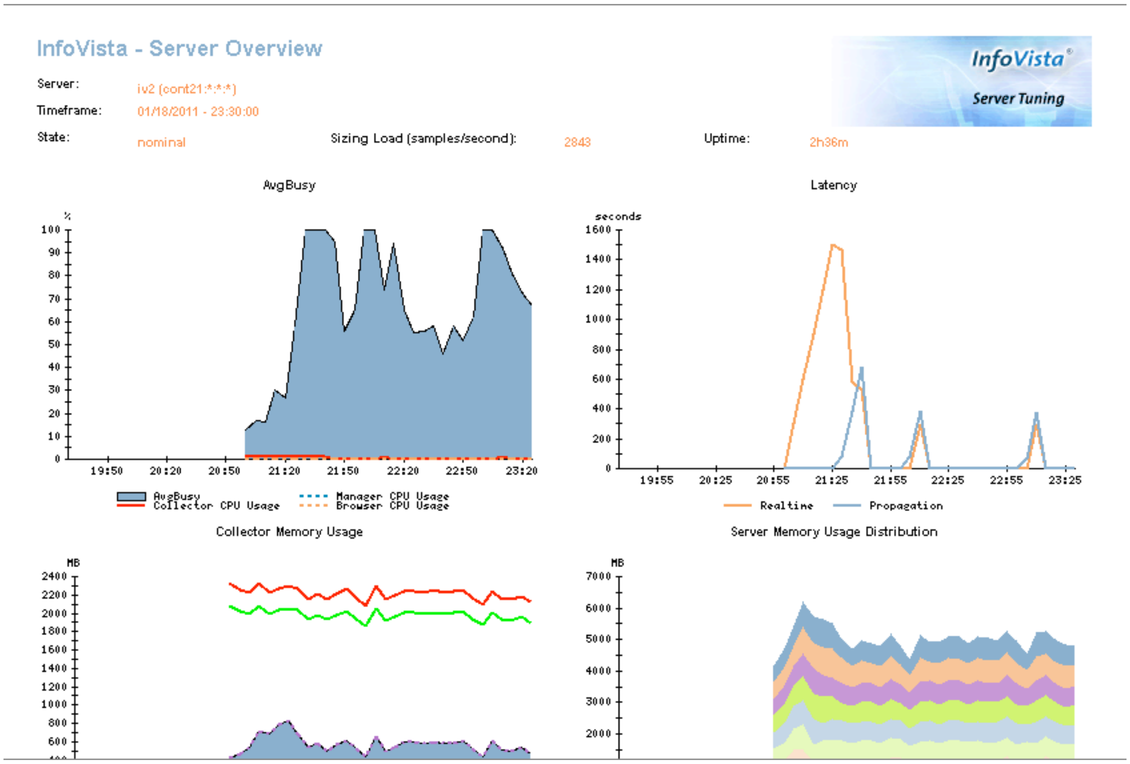
Key Performance Indicators Used for InfoVista Server

Different operating system and storage metrics were collected during the tests, yet the most important KPI was related to the application itself and used for checking the health of an InfoVista Server: `avgBusy` is a counter provided by the tuning VistaView, which is a built-in library that provides health and performance indicators of the InfoVista Server itself. The `avgBusy` counter reflects the run-queue health of the processor.

An `avgBusy` of 70% is considered as a critical threshold above which InfoVista Server starts to postpone processing the data it collects in real time: the InfoVista backlog increases, the propagation tasks (tasks with less priority than real-time pooling) ends up in wait mode and the propagation backlog increases. If the avgBusy stays above 70%, InfoVista Server may not be able to monitor the network devices anymore.

In production environments, sizing and monitoring techniques are put in place to prevent going up to a crash when the 70% threshold is crossed.

The figure below provides a screen shot of the tuning VistaView in an InfoVista Server:



The I/O profiling was performed with iostat and zpool iostat on the T5240, and with the Analytics tool of the Sun ZFS Storage Appliance for the SAN.

Baseline and Stress Scenarios

Phase 1: A baseline was established to measure improvement during the rest of the test campaign. The baseline configuration consisted of:

- One Solaris Container hosting 8 InfoVista Server instances and a single ObjectStore instance
- No separate ZFS pool for the database transaction logs and no SSD. A single ZFS pool was shared between the transaction logs and the data thereby absorbing all the I/O operations
- Default ZFS settings. The ZFS properties were set as out-of-the-box. Especially the recordsize was equal to 128K, and the size of in-memory for data and metadata limited to 4GB

This baseline reproduced the behavior of InfoVista Server and its local database without any virtualization and without any Oracle Solaris ZFS or SSD optimization. The baseline was stressed with an average workload that went up to 8 x 2500 samples/sec generated by a topology of 80K instances.

Phase 2 : Once the baseline established, a second phase consisted in running and monitoring the following stress scenarios in order to improve the storage performance. During this second phase, only one Oracle Solaris Container (i.e. 8 InfoVista Server instances) was used:

1. Baseline plus changing the ZFS `recordsize` to 4K and then to 8K
2. Baseline plus:
   - `ZFS recordsize=4K`
   - 2 striped SSD for the ZFS Intent Log of the unique ZFS pool
   - 3. Baseline plus:
   - `ZFS recordsize=4K`
   - A separate ZFS pool on 2 striped SSD for the ObjectStore transaction logs
   - No SSD for the data ZFS pool
   - 4. Baseline plus:
   - `ZFS recordsize=4K`
   - A separate ZFS pool on 2 striped SSD for the ObjectStore transaction logs
   - 2 striped SSD for the ZFS Intent Log of the data ZFS pool
   - 5. Baseline plus:
   - `ZFS recordsize=4K`
   - A separate ZFS pool with 2 striped SSD for the ObjectStore transaction logs
   - 2 striped SSD for the ZFS Intent Log of the data ZFS pool
   - 2 striped SSD for L2ARC of the data ZFS pool

The latest configuration - #5 - defined an optimum in terms of storage performance.

Phase 3: During a third phase this optimum configuration was tested against three Oracle Solaris Containers, hence running 24 InfoVista Server instances and three ObjectStore instances on the Sun SPARC Enterprise T5240. Each container sustained a workload of 8 x 2500 samples/sec for a total of 60K samples/sec generated by a topology of 240K instances.

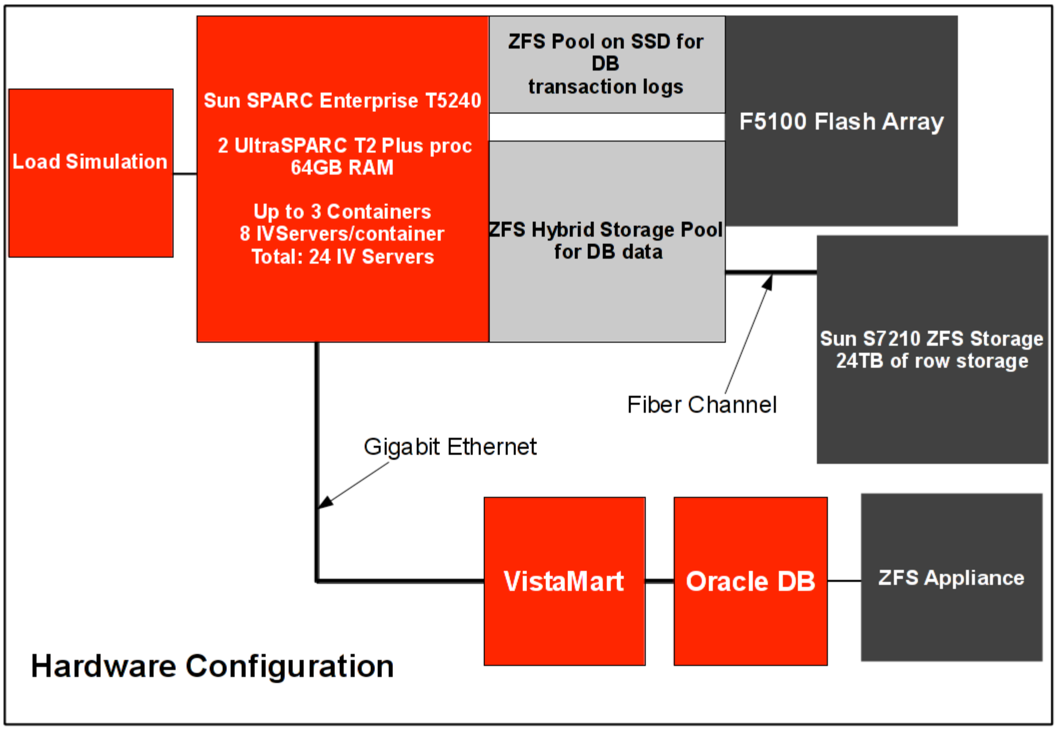
Test Environment

Hardware Configuration

In order to reproduce a real customer environment the configuration was not limited to the T5240 server, the Sun Storage 5100 Flash Array, and the Sun ZFS Storage 7210 Appliance. It also included hardware required to deploy the VistaMart component and its associated central database. A second Sun ZFS Storage Appliance was dedicated to the central database.

The hardware and storage configurations for VistaMart and the Oracle database have not been optimized and for this reason are not described in details. They were sized large enough for these components not to act as a bottleneck during the test.

Finally, a separate server was used to generate the workload and simulate the network devices.



Storage Configuration

The two Oracle Solaris ZFS pools used on the Sun SPARC Enterprise T5240 during the test were configured as follow:

- One pool dedicated to the ObjectStore database transaction logs and consisting of 2 striped SSDs from the F5100 Flash Array (a third striped SSD was later added to this pool, but no performance increase was observed)
- One pool dedicated to the ObjectStore data using:
- The 42 SATA HDD of the Sun S7210 ZFS Storage
- 2 striped SSDs from the F5100 Flash Array for the ZIL
- 2 striped SSDs from the F5100 Flash Array for the L2ARC (ZIL and L2ARC were also added a third striped SSD, but without any performance increase)

The following figures show the configuration details:

```
pool: objStore
 state: ONLINE
 scrub: none requested
config:

        NAME                       STATE     READ WRITE CKSUM
        objStore                   ONLINE       0     0     0
          c3t2100001B329A045Ed0    ONLINE       0     0     0
        logs
          c4t1d0                   ONLINE       0     0     0
          c4t2d0                   ONLINE       0     0     0
          c4t10d0                  ONLINE       0     0     0
        cache
          c4t5d0                   ONLINE       0     0     0
          c4t6d0                   ONLINE       0     0     0
          c4t11d0                  ONLINE       0     0     0

errors: No known data errors

  pool: transactionLog
 state: ONLINE
 scrub: none requested
config:

        NAME            STATE     READ WRITE CKSUM
        transactionLog   ONLINE       0     0     0
          c4t3d0        ONLINE       0     0     0
          c4t4d0        ONLINE       0     0     0
          c4t9d0        ONLINE       0     0     0
```

```
NAME                         USED  AVAIL  REFER  MOUNTPOINT
objStore                     752G  3.66T    23K  /objStore
objStore/cont21              376G  3.66T   250G  /objStore/cont21
objStore/cont22             18.3G  3.66T  18.3G  /objStore/cont22
objStore/cont23             71.1G  3.66T  71.1G  /objStore/cont23
transactionLog              4.43G  63.1G   112K  /transactionLog
transactionLog/cont21       1.83G  63.1G  1.04G  /transactionLog/cont21
transactionLog/cont22       1.18G  63.1G  1.18G  /transactionLog/cont22
transactionLog/cont23       1.39G  63.1G  1.39G  /transactionLog/cont23
```

Performance and Scalability Results

Before the test campaign, the maximum performance achieved by VIN on a single Sun server was 12.8K samples/second. This rate was achieved by running 8 InfoVista server instances each running at 1.6K samples/s equating for a total of 12.8K samples/s per physical server.

After the test was completed, the performance achieved on the same physical hardware showed an increase of 3.75x. This large gain can be attributed to:

- InfoVista's platform multi-process architecture and,
- Oracle's virtualization and storage technologies, which powered the creation of 3 containers. Each container running 8 InfoVista servers at a rate of 2K samples/s.

This parallel processing capability combined for a total rate of 48K samples/s -3.75x more than before the study.

Best Practices

Storage Tuning

Following are best practices in terms of storage tuning to improve IV Server performance:

- Putting the ObjectStore transaction log on a very fast storage compared to the storage for the ObjectStore data does not suffice to bring a lot of performance improvement: the transaction log fills in too quickly and the storage pool dedicated to the data can not keep up with pace. Thereby the optimal configuration consists in putting the transaction log on a dedicated storage pool made of 2 striped SSDs and putting the data on an Hydrid Storage Pool that also benefits from 2 striped SSDs for its ZIL and L2ARC. Such a configuration allows load performance to be doubled
- In addition, Striped SSDs as ZIL strongly reduce the IOPS of the data and allow for the database to be put on entry level SAN (with slow SATA hard disks)
- Striped SSDs for the Hydrid Storage Pool L2ARC allow a better read of the performance of IV Server and simplify the recovery and push process of IV Server (Read flow for VistaMart)
- Changing the ZFS record size from 128K to 4K dramatically reduces the latency thanks to better disk block alignment

Virtualization And Sizing

Oracle Solaris Containers allow scaling of IV Server:

- Deploy Oracle Solaris Containers for IV Server on highly multi-threaded T-Series servers
- The optimal configuration for a container is 8 IV Server instances and 1 ObjectStore instance
- From there, the optimal configuration for a T5240 with 64GB of RAM is 3 containers for a total of 24 IV Server instances

Summary

Many interesting conclusions can be drawn from this study.

First, scalability was achieved without modifying the application code: the advent of multi-core CPU as a de-facto industry standard does not mean that applications need to be redesigned and rewritten with the objective of using more threads. In the case of VistaInsight for Networks, virtualization was used to multiply the number of IV Server instances running concurrently on a single system. Oracle Solaris Container technology fulfilled its promise of a very low virtualization overhead: verification has been made that the application performance within a single container is equal to that of the application running directly on the operating system—i.e. without virtualization.

Second, high application throughput performance and scalability required not only good CPU but also good I/O performance. For a real world solution that comes not only with a business logic layer but also with a database layer, the overall performance cannot be reduced to a single metric such as CPU performance. From that perspective, VistaInsight for Networks is a typical case where performance required a balanced hardware infrastructure capable of spreading its CPU and I/O power between many application instances. The adjustment of CPU and I/O capacities to the application needs—such as dedicating SSDs to critical transaction log files—represented the major part of the improvements achieved during the test.

As a result, the solution performance improvement came from the fine integration between its different components: the application, the virtualization layer, and the hardware platform. This integration was made possible by the flexibility that both the Oracle Solaris Container and the ZFS Hybrid Storage pool offer.

InfoVista's and Oracle's customers can now benefit from the work and the results achieved during this study. Oracle Solaris Container, Oracle Solaris ZFS, and Sun ZFS Storage Appliance are now fully supported by InfoVista.

Scalability benefits are offered through Oracle Solaris Container technology that maximizes the resource usage of already deployed hardware.

Performance increases are ensured through intelligent usage of ZFS and SSDs: more load can be supported by each single polling instance, whereas the storage performance is reduced thanks to the caching mechanisms of ZFS and their combination with SSDs.

Appendix

APPENDIX – 1: TEST CONFIGURATIONS AND RESULTS

Scenario	S7210 IOPs	S7210 Latency	CPU	samples/sec	AvgBusy()	max containers	max IVServers	Remarks
Base line 1 IVServer system with 8 IV instances: * one ZFS pool, one ZFS file system * no SSD * ZFS default settings (rec size=128K, ARC for data & metadata limited to 4 GB: set zfs:zfs_arc_max = 4294967296) **S7210:** * no SSD * one mirrored pool, one LUN with rec size=8K	average : 948 max: 30544	average : 13 ms max: 27 ms	average usr+sys: 2% max usr+sys: 63%	roughly 8 x 2500	80%	1	8	32 bit of IVServer process load must be kept under 3000 samples/sec AvgBusy() is above 60 % and will overload in case of recovery or GC
Base line + 1 IVServer system with 8 IV instances: * ZFS rec size=4K for ObjStore data & ObjStore transaction log	average : 737 max: 27669	average : 4 ms max: 10 ms	average usr+sys: 2% max usr+sys: 42%	roughly 8 x 2500	75%	1	8	AvgBusy() is above 60 % and will overload in case of recovery or GC
Base line + 1 IVServer system with 8 IV instances: * ZFS rec size=8K for ObjStore data & ObjStore transaction log	average : 1001 max: 29459	average : 5 ms max: 22 ms	average usr+sys: 3% max usr+sys: 57%	roughly 8 x 2500	75%	1	8	AvgBusy() is above 60 % and will overload in case of recovery or GC
Base line + 1 IVServer system with 8 IV instances: * 2 striped SSDs for ZIL of ObjStore data and ObjStore transaction log, SSD from F5100 direct attached * ZFS rec size=4K for ObjStore data and ObjStore	average : 624 max: 30338	average : 4 ms max: 12 ms	average usr+sys: 4% max usr+sys: 39%	roughly 8 x 2500	65%	1	8	AvgBusy() is above 60 % and will overload in case of recovery or GC
Base line + 1 IVServer system with 8 IV instances: * 2 striped SSDs for ObjStore transaction log, SSD from F5100 direct attached * no ZIL, no L2ARC * ZFS rec size=4K for ObjStore transaction log	average : 865 max: 37930	average : 8 ms max: 51 ms	average usr+sys: 5% max usr+sys: 75%	roughly 8 x 2500	60%	1	8	transaction log file went to 2 GB, and IVS crashed
Base line + 1 IVServer system with 8 IV instances: * 2 striped SSDs for ObjStore transaction log, SSD from F5100 direct attached * 2 striped SSDs for ZIL, no L2ARC * ZFS rec size=4K for ObjStore transaction log	average : 604 max: 25393	average : 4 ms max: 8 ms	average usr+sys: 3% max usr+sys: 45%	roughly 8 x 2500	50%	1	8	Ivs performances increased, yet additional L2ARC would help the push flow (from IVS to VM)
Base line + 1 IVServer system with 8 IV instances: * 2 striped SSDs for ObjStore transaction log, SSD from F5100 direct attached * 2 striped SSDs for ZIL, 2 striped SSDs for L2ARC * ZFS rec size=4K for ObjStore transaction log	average : 507 max: 1919	average : 11 ms max: 30 ms	average usr+sys: 5% max usr+sys: 57%	roughly 8 x 2500	50%	1	8	IOPS are dramatically reduced on S7210
3 times Base line + 3 containers with each 1 IVServer system with 8 IV instances, for a total of 24 IV instances: * 2 striped SSDs for ObjStore transaction log, SSD from F5100 direct attached * 2 striped SSDs for ZIL, 2 striped SSDs for L2ARC	average : 1885 max: 27178	average : 12 ms max: 15 ms	average usr+sys: 12% max usr+sys: 59%	roughly 3 x 8 x 2500	65%	3	24	Number of IOPS averages around 2000 with peaks close to 30000

Note: Since this test was run, the S7210 has been replaced in the Oracle storage catalog by the [Sun ZFS 7120 Appliance]

Reference Architecture For VistaInsight for Networks on Oracle Server And Storage

Author: Thierry Manfe, Michel Schuster
December 2011 Version 1.0

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506. 7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Copyright © 2011, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 1010

Hardware and Software, **Engineered to Work Together**