NexentaStor Enterprise

Backend for CLOUD

Marek Lubinski

Sr VMware/Storage Engineer, LeaseWeb B.V.

AGENDA

- LeaseWeb overview
- Express Cloud platform
- Initial storage build
- Why NexentaStor
- NexentaStor initial design/sizing
- Base migration and challenges
- What went wrong
- Sizing review, changes
- Final configuration, what next
- Lessons learned

About LeaseWeb

- Founded in 1997
- Part of the OCOM group
  - DataXenter (Modular Data Center Technology)
  - EvoSwitch (Carrier-neutral Data Center Services)
  - Fiberring (Network Services)
  - LeaseWeb (Hosting Provider)
- Cloud, Dedicated Servers, Colocation, Hybrid Infrastructure
- >50.000 physical servers in 6 data centers
- >3,0 Tbps bandwidth capacity
- IaaS Cloud services: Private/Hybrid Cloud, Premium Cloud, Express Cloud

Express Cloud platform

- **Old situation:**
  - VMware vSphere 4.x clusters
  - Mixed hardware vendors
  - Storage: Supermicro & Dell R510 (Linux NFS)
- **New situation:**
  - Cloudstack backend with KVM hypervisors
  - Standard hardware: Dell, HP
  - Storage: NexentaStor Enterprise (NFS)

Initial storage build

- **Supermicro: 24x SATA in RAID10, 1Gbit, Linux NFS**
- **Design challenges:**
  - Performance issues (50-200 VM's per box)
  - SPOFs
  - rack space
  - Scalability
  - Flexibility
  - Expensive growth
  - Lack of performance indicators.
- **Thousands of VMs running**

Why NexentaStor

- **ZFS**
  - Resiliency
  - Scalability
  - Performance
  - Snapshots
  - Compression
- **Native NFS (also CIFS+iSCSI support)**
- **Dtrace**
- **HA cluster (Active/Passive or Active/Active)**
- **no HW vendor lock-in (heads, JBODs, components, disks)**

NexentaStor initial design/sizing

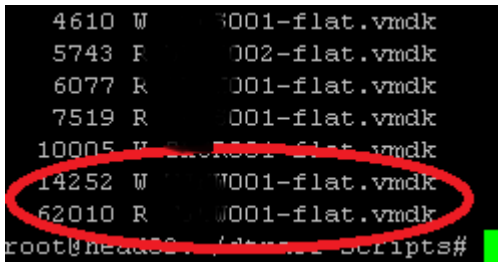
- **Estimations:**
  - Platform IOPs -> max. 40k IOPs per setup
  - Avg IO size: 4-8k
  - Read/write ratio: 70/30
  - Cache/hit ratio (when compared to a different platform with PAM2): 30%
- **Design/configuration:**
  - HA cluster (Active/Passive)
  - RAIDZ-2 (11vdevs, 6disk/vdev)
  - NL-SAS drives +2x ZeusRAM SSD 8GB (ZIL) + 4x OCZ SSD (L2ARC), 3x JBOD
  - Supermicro head: 2 CPUs QC + 96GB RAM, LSI SAS HBA's, 10Gbit Intel NICs
  - One rack
  - Network backend: EX4200 Juniper switches + Cisco 3750X, 10Gbit to storage

Base migration and challenges

- **Start of Cloudstack POD setup on Nexenta storage**
- **Migration of vSphere workloads:**
  - Scripted storage VMotion
  - Problems with storage VMotion after ~1200 VM's migrated (NFS kernel tweaks)
- **Performance:**
  - Iometer benchmark

Test name	Latency	Avg iops	Avg MBps
Max Throughput-100%Read	24.25	2477	77
RealLife-60%Rand-65%Read	14.91	3048	23
Max Throughput-50%Read	19.06	3012	94
Random-8k-70%Read	5.28	6442	50

- **Rapid growth causing performance issues:**
  - Iometer low results
  - Increased latency to 100msec
  - Maxing ZIL throughput
- **Tracing storage abusers:**
  - Swapping ---------->

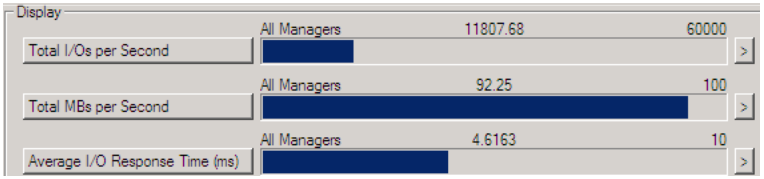


What went wrong

- **Bigger average IO size**
  - 8-32k
  - It maxed estimated 40K IOPs
- **Different read/write ratio**
  - Read/write: 20/80
  - 90% random
- **NFS sessions default limit**
  - After 1200 VM's running storage vMotion started to fail
  - No new disk IO operations were possible (copy, etc.)
  - Current workloads were not affected
- **NFS timeout settings recommendations**
  - esxcfg-advcfg -s 24 /NFS/HeartbeatFrequency
  - esxcfg-advcfg -s 10 /NFS/HeartbeatTimeout
- **Large growth of CS VMs caused high peaks in NFS Ops resulting in increased latency to data stores**

Sizing review, changes, data migration

- **New VM Pool**
  - Only mirrored VDEVs using only SAS15k drives
  - More ZeusRAM disks (ZIL)
  - More OCZ SSD disks (L2ARC)
  - More memory (192G) - ARC
  - Bigger heads (more LSI SAS cards)
- **Iometer benchmark**
  - 100% random
  - 70% read



- **Old data migration (scripted VMware storage vMotion)**
- **Cloudstack migration plan (ZFS sync + NFS remount)**

Final configuration and next steps

- **Desired configuration**
  - 2x VM Pool on SAS15k
  - Multiple ZeusRAM (ZIL)
  - Multiple OCZ SSD (L2ARC)
  - In case of performance issues option to go active/active setup
  - Very scalable with decent performance
- **Future steps:**
  - Exact build for other locations (Frankfurt, U.S.)
  - Off-site replication for data protection

Lessons learned

- Use only mirrored VDEVs with high expectations
- Achieved cache/hit ratio: 90%
- Use Dtrace to find storage anomalies (abusers)
- Avoid workarounds (they become permanent)
- Play around with the system in test environments
- “Nexenta architect training” highly recommended

THANK YOU