Real world flash practices - the Cisco point of view
16th TERENA TF-Storage at ACONet in Vienna, Austria

Alexander Weinbacher
aweinbac@cisco.com
Feb 2015
Cisco does servers!

...and storage, too!
Agenda part 1

• Storage Performance *(and how you can avoid it easily)*
  • Metrics *(and what they don’t tell you)*
  • Real-world numbers *(and how to cook them)*
  • Most common pitfalls *(including my own favourites)*

• All flash array basics
  • Flash NAND and media: SLC vs MLC vs TLC
  • AFA vs traditional scale-up architecture
  • Advantages and limitations
Agenda part 2

- Cisco Unified Computing System
  - Invicta solid state storage
    - Performance reports
  - UCS architecture
  - UCS service profiles
  - Use cases for Invicta
Part 1

Storage performance

(and how you can avoid it easily)
Why storage performance?

• 100% in-memory environments are still rare. We will still need fast external arrays for some years.

• 70% of all performance problems have the same root causes, I see the same mistakes happening over and over again!

• These common mistakes are easy to correct for the knowing.
But what is “fast storage”?

- Talking about performance, we first need consensus on:
  - Metrics and units: IOPS, microsec, MB/s, SPC-01/02
  - Relevance of metric for my situation: Backup, OLTP, DSS, SMB/NFS fileshare?
- Test procedure
- Test tools (IOmeter, LoadDynamix, Finisar, tools on the array)
- Test interval (which is the relevant period?)
- Interpretation / benchmark of the results
Metrics for storage performance - KPI

- The key performance indicators and what they stand for
  - Throughput in MB/s
    - the ability to move data from one place to another
  - Input-Output per second in IOPS
    - scalability, always state the block size!
  - Latency in us/ms
    - the KPI for database storage. But the least understood.
IOPS vs latency on a single FC 15k disk
Never trust a single metric!

- It is the combination of metrics that tells the truth!

Source: [http://searchstorage.techtarget.com/NetAppSponsoredNews/New-Benchmark-Results-for-Unified-Scale-Out-Storage](http://searchstorage.techtarget.com/NetAppSponsoredNews/New-Benchmark-Results-for-Unified-Scale-Out-Storage)
Do the math! - IOPS vs. latency

- The connection between IOPS and latency
  - 1
  - ------------------------ = 1 IO useable for the application
    tsan + t io-operation

- tsan = latency of the SAN (s)
- t io-operation = time to handle the IO without the SAN (OS-driver + HBA, cabling) (s)
SAN latency in real-world scenarios

• Example based on older gen. MDS9500 model whitepaper:

• Figure 2 shows that observed latencies varied greatly, frequently reaching peaks of 36 microseconds or higher, almost 10 times what is promised in the vendor's marketing collateral. By contrast, Cisco MDS 9500 Series latency was observed to remain consistent at 13 to 15 microseconds across all ports as loads increased.

SAN latency in real-world scenarios

**Netapp FAS:** Example output from the *stats show* command placed in a table for ease of display.

<table>
<thead>
<tr>
<th>Instance</th>
<th>avg_latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>28.12</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>9.36</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>11.38</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>28.27</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>22.88</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>11.75</td>
</tr>
<tr>
<td>/vol/volume_name/lun1-Xn/8W3OdfJbw</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Synchronous mirrored arrays with FC-SAN directors and HDDs

The additional SAN-latency is negligible with traditional HDD-based arrays!
Now let’s do the Math again!

- HDD-array with 10ms LUN latency
  - 13us SAN latency, 4 hubs – leads to 0.52% loss of IO performance
    \[
    \frac{1}{(4 \times 13 \times 10^{-6} + 10 \times 10^{-3})}
    \]

- SSD-array with 150us LUN latency
  - 13us SAN latency, 4 hubs – leads to 25.74% loss of IO performance!!!
    \[
    \frac{1}{(4 \times 13 \times 10^{-6} + 150 \times 10^{-3})}
    \]

- Conclusion: The faster the array works, the more relevant is the SAN latency and the cabling length!
But how can we solve that latency problem?

- The less components between disk and server, the better
- Physics tell you: Get rid of that SAN!
- IT operation team tells you: I have to make my life easier, not more complex! I need my SAN!
- PCI flash inside the servers are a quick fix, but not an economic one and a very risky one, too (lack of proper RAID!)
- Infiniband is fast, but administration not easy like SAN and you would end up with a third network (LAN, SAN, IB)
- Storage vendor tells you: Pimp your array! More cache, buy SSDs!
Be careful with fancy features!

- Test results from first generation of EMC FAST tiering feature: IOPS have increased, but write latency doubled!
- Can be negligible in some environments (Exchange), but dramatically in others (OLTP)!
- Source: https://communities.netapp.com/community/netapp-blogs/shadeofblue/blog/2011/01/19/fast-is-slow-itsquo-s-the-latency
...and even in the current version!

- More IOPS, but also more latency!

One solution: All flash arrays

- Offer considerably more CPU-power per TB than traditional arrays
- Are designed for a few us disk-latency
- Inline deduplication/compression
- Hybrid systems have diskspace-tiering between SSD and SATA
- Even under load, they deliver latencies of only 150us-1ms!!!
Do not watch unless willing to be confronted with datacenter reality!

The content herein might be shocking for some of you because it is too simple to believe that it happens every day!

It is the essence from all the mistakes and dramas I have seen in my 15 years+ of storage work, some of those even caused by myself!

If you follow this advice, you will avoid 80% of common mistakes and have enough time left to invent and add some of your own 😊
Most common pitfalls for bad performance

- HDD overload and/or large IO-queues that slow the server down
- Ignoring the FC-Interopmatrix: Firmware versions don’t fit
- Server timeout-settings not according to array and server OS best practices (very important for failover scenarios!)
- LUN alignment not corrected for the OS
- Primary FC-path assigned wrong (data forced to flow over storage controller-interlink)
- Backend-cabling of array faulty or slowly dying SFPs
- Design mistakes: Many don’t know which metrics are relevant for their situation and which numbers to ask for in the RFP.
Beware of onboard tools!

• Onboard storage array tools are mostly inaccurate
• Treat their results only as a rough indication, not as a precise measurement!
• Most of them only measure a 1-minute average or maybe a 5-10 sec-average!
• If you have performance peaks below 1 sec intervals, which I have seen often, you will not find them within those onboard tools!!!
• That will lead to improperly sized arrays and an unhappy customer!
• Do not use good ole´ IOmeter for load testing! It has proven to act inconsistently, use more accurate tools like LoadDynamix etc.
Part 1

All flash array basics
NAND Flash

- Writes only full 1MB / 2MB blocks
- First erase, then write = write amplification
- Every write hurts lifetime
- Avoid unnecessary writes
- You need a decent write optimization!

Symmetrical read-write optimization

• How Cisco Invicta does it – "Block Translation Layer"

Inbound data blocks

Fill

Write

Cache

Block Translation Layer

Write
Flash media types

<table>
<thead>
<tr>
<th>type</th>
<th>writes</th>
<th>speed</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC</td>
<td>10.000</td>
<td>++++</td>
<td>$$ $$</td>
</tr>
<tr>
<td>eMLC</td>
<td>5.000</td>
<td>+++</td>
<td>$</td>
</tr>
<tr>
<td>MLC</td>
<td>3.000</td>
<td>+++</td>
<td>$</td>
</tr>
<tr>
<td>TLC</td>
<td>1.000</td>
<td>++</td>
<td>$</td>
</tr>
</tbody>
</table>

• But: Latest TLCs offer better consistent performance than recent generations and you can use so many more than MLCs!
• Reliability of SSD vs HDD?
The revolution of storage – All Flash Arrays

- No HDDs anymore
- Only SSDs
- 100-1000x faster than HDD
- More reliable than HDD
- High CPU power
- Scale-up and Scale-out
- Many not yet enterprise-grade: NDU, SW features
- Cisco Invicta series is a true AFA
Hyperconvergence – key driver for flash

Bladeservers for BareMetal-OS or Hypervisor (VMWare, Hyper-V)

Storage Controllers = Intelligence and Management (Fault Tolerance, High Availability, Mirroring, Clustering, Snapshots)
Cost of SSD vs HDD?

- SSDs more expensive than HDDs?
- Compression / dedupe inline on SSD, post process on HDD
- Rackspace, power
- SSD AFA vs. local SSD caching disks?
Part 2

Cisco Unified Computing System
Traditional hard disks are not dead, but......
Cisco UCS Invicta Series

UCS Invicta Scaling System

UCS Invicta Appliance

FIRST RELEASE
235,000 IOPS
1.3 GB/s Bandwidth
Up to 24 TB RAW
Max. 1ms latency

FIRST RELEASE
Up to 1.3 Million IOPS
Up to 13.2 GB/s Bandwidth
Up to 240TB RAW

RAID 6, 144TB raw
FC, iSCSI (NFS soon)
Inline Deduplication
Thin Provisioning
Snapshots
Array-based replication

© 2013-2014 Cisco and/or its affiliates. All rights reserved.
Invicta ORION Advanced OLTP performance

24 SSDs, 1 Controller, (smallest building block)

70us!!!
Invicta ORION Advanced OLTP performance
100 VDI desktops boot in 70 sec!
Invicta use cases

- Virtual Desktops
- Transcoding
- OLTP
- ERP
- Database Loads
- Batch/Month End Processing
- Analytics & Intelligence
Managing Compute and Storage in the Unified Fabric

UCS Central | UCS Manager | UCS Director

Fabric Interconnects

Switched Fabrics

Stateless UCS Servers
With Virtualized Adapters
Unified Computing System

- Computing, Networking, Storage united
- Service Profiles are object-based approach
- Service Profiles contain config details, independent from HW
- Upgrades and downgrades via Service Profile (is an xml file)
- No matter if blade or rackmount, as long as on Fabric Interconnect
- Always the same passive chassis
- FI is the closest way to connect external storage – lowest latency
Service Profile Templates + Storage Profiles
A Powerful Combination

Service Profile Template 1

Service Profile 1

Boot LUN: 80GB
Data LUN1: 20GB
Data LUN2: 10 GB

Service Profile 2

Boot LUN: 80GB
Data LUN1: 20GB
Data LUN2: 10GB

Service Profile 3

Boot LUN: 80GB
Data LUN1: 20GB
Data LUN2: 10GB

Service Profile 4

Boot LUN: 80GB
Data LUN1: 20GB
Data LUN2: 10GB

Create Storage Profile
Create Service Profile Template and Consume Storage Profile
Create Service Profiles from Service Profile Template
Associate Service Profiles to Servers

LUNs Automatically Provisioned and Masked on Storage Blade
A Powerful Combination
Next Wave of Unified Computing
Powering Applications at Every Scale

Fourth Generation UCS Servers

UCS M-Series Modular Servers
UCS C3160

Edge-Scale Computing
Core Data Center Workloads
Cloud-Scale Computing

UCS Management
Compute • Network • Storage • Virtualization • Management and Automation

UCS Mini

Core Data Center

UCS Management Innovation
Did you like it?

• YES: Thanks and please recommend it to your friends!

• NO: Thanks and recommend it also!

(Because your enemies should also waste their time!)

😊
TOMORROW starts here.