Education

Operating System, Storage Performance Analysis

Robert M. Smith, Microsoft Corporation

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Abstract

OS Storage Performance Analysis

- Analyzing and dealing with storage performance at the OS level can be challenging in many respects. This tutorial covers aspects of performance with respect to storage. This tutorial will also cover tools that can be used to assist in the analysis of operating system performance.

Tiered Storage Architecture

Education SNIA

Tier-0

- Ultra-Performance
- Time-sensitive, transaction-based
- SSD or Hybrid (SSD + traditional)

Tier-1

- "Performance Optimized" drives
  - Highest RPM (10k, 15k)
  - <=600 GB
  - 3.5" or 2.5"
  - Fibre

Tier-2

- "Capacity Optimized" larger drives
  - Lower RPM (5400 or 7200)
  - 500 GB – 3 TB

Tier-3

- 2 TB, 3TB, up to 6 Gb/s interface
- Tape
- Optical
- Archival
- Long-Term

"Rotational" Disk Drives

- "Rotational" disks:
  - Cost: Pennies per GB (not including enterprise costs)
  - "Capacity Optimized" drives
    - TB Size: 0.5, 1, 2, 3
    - IOPS: >= 180 (depending on I/O profile)
      <Random workloads>
    - SAS or SATA
    - Regardless of size, same performance, same IOPs
    - 4-8 ms latency (on average across manufacturers)
  - "Performance Optimized" drives
    - GB Size: 75, 150, 300, 500, 600
    - IOPS: 200-250 or so (worst case)
    - SAS, FC (some SATA)
    - 2-4 ms latency (on average across manufacturers)

Solid-State & Hybrid Storage

Education SNIA

- SSD & Hybrid Storage
  - Cost: Dollars per GB
  - SSD Solid-State Drive
    - No moving parts
    - Less power consumption
    - 75, 150, 300, 500, 600 GB
    - OS likely has native SSD support (Trim, etc)
    - Microsecond latency
    - Flash block erase before write
  - Hybrid Storage Solutions
    - Solid-State and rotational disks in same chassis
    - "Hot" data serviced by SSD, other serviced by rotational

Storage Hardware Configurations

Education SNIA

Controller Cache Configurations

- How much cache?
- What is read/write ratio of cache?
- How effective is cache?
  - Enterprise storage usually has performance measuring capability onboard
- What happens when a threshold is reached? (I.E. Flush)
  - **Idle flushing:** does not interrupt, I/O continues
  - **Low and high watermark flushing:** triggers flushing, minor performance impact
  - **Forced flushing:** to free cache pages, all I/O temporarily halted

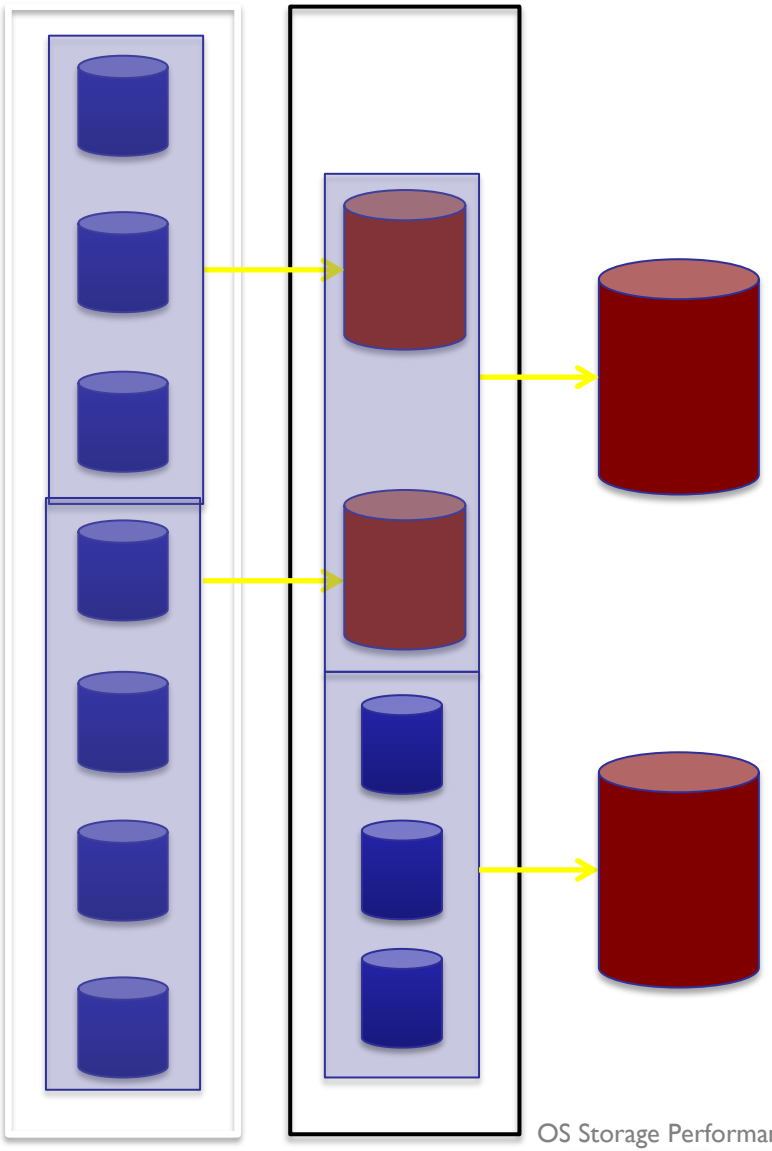
Storage Hardware Configurations - 2

- Is cache "mirroring" involved
  - If so, is there a performance impact?
- Are there other workloads on the storage device?
- What hardware is in between the initiator and target?
  - If SAN, how many and what types of switches?
  - Virtualization Appliances
    - Some take the "LUNs" presented and virtualize those
    - Some have onboard storage

Storage Hardware Virtualization

- What exactly is a Logical Disk?
  - Comprised of a group of physical disks
  - Slice or dice of a "pool" of physical disks
  - Aggregation of underlying LUNs (virtualization appliance)
- What indicates a logical disk?
  - Non "standard" size (150, 300, 500, 600, 1 TB, 2 TB)
  - Use OS tools to interrogate hardware
  - Use vendor tools to interrogate hardware

Virtualization Appliance Example



Reasons:

- Pool storage; existing and new increases efficiency
- Management efficiency

Issues:

- Additional Complexity
  - Management
  - Troubleshooting
- More ports to manage
- More potential congestion points

Storage Layout

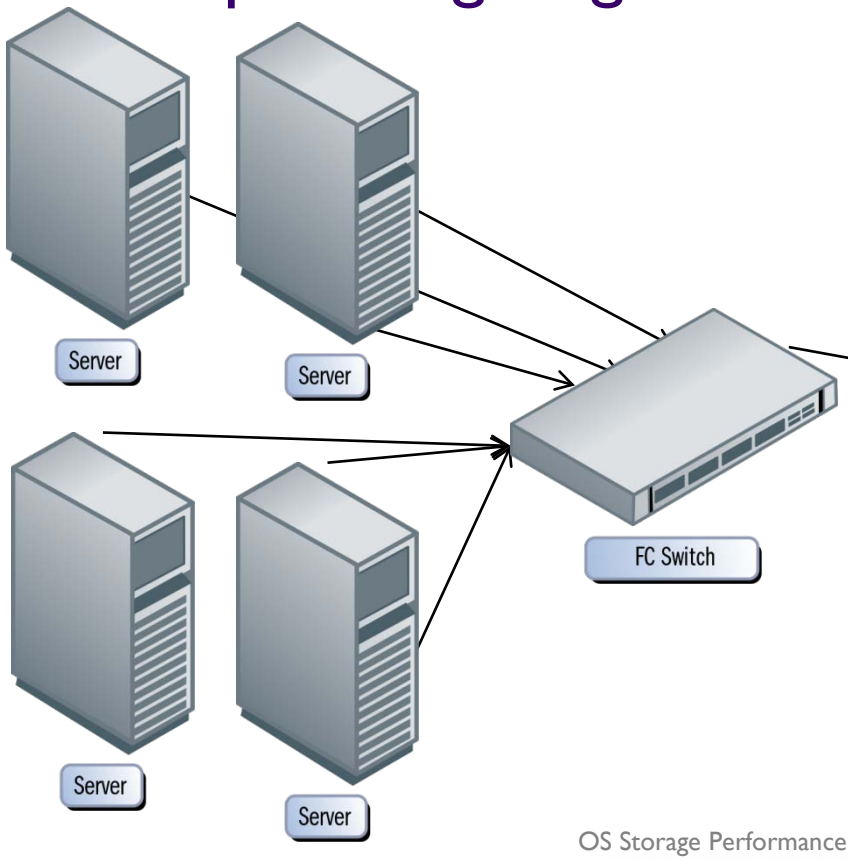
- How are the disks configured?
  - RAID level (ex. 5, 6, 1+0, 5+0, etc)
  - Number of physical disk drives backing
  - Levels of virtualization between server(s) and disks?
  - Any storage pool sharing involved?
    - Dedicated disks or shared storage pools?
  - What is the backup schedule (I.E. any "invisible I/O"?)
  - Who was involved in design? (purchasing, architects?)
- What happens to the pool if a disk drive fails?
  - What is the performance impact?
  - How long to rebuild?
  - Data could be vulnerable during rebuild

Path(s) to storage

- Storage Area Network (Fibre Channel & iSCSI)
  - Path is usually a "mesh"
  - Redundant paths are common
  - Multipath I/O (MPIO) software can load balance
  - Paths may have different loading based on design (or lack thereof)
    - Oversubscription
  - There may be a number of intermediate devices
    - Core switches
    - Routing could be a factor
    - Inter-switch links (ISLs) can be a factor

SAN Congestion

- Many ports do not guarantee performance
- Common ratio is 5:1, host ports to target ports
- Keep a living diagram of SAN to avoid congestion



Mass-Storage Controllers

Education SNIA

Mass-Storage Controllers

- Range from simple on-board to complex, battery-backed, RAID
- Basic controllers report limited diagnostic information
- Advanced controllers have diagnostics available
  - Vendor supplied tools
  - Capable of sending events to operating system through extended logging

Host-Bus Adapters

- SCSI command interface to OS
- Often synonymous with Fibre Channel SAN
- Offload packet assembly and disassembly
- Provides OS a view into the SAN
  (though most is abstracted by default)
- Mainly Fibre Channel or SAS

Network Interface Cards (NIC)

- Some NICs have offload capability
- Teaming software for speed, throughput and availability
- 10 Gb/s common
- Network Monitor and Wireshark available unless using IPSEC

Latency

- Rotational Disks
  - Millisecond latency
  - Sequential writing to rotational drives is the most efficient
  - Sequential, and/or "full-stripe" writes to RAID disks are most efficient
  - Latency occurs as actuators have to move to physical location
  - Operating system logical address may be different from physical location on disk device
- SSD
  - Microsecond latency
  - Small random writes slowest (Flash block)
  - Flushing
  - Firmware
    - Has improved performance significantly

Queuing

- The art of keeping the I/O pipeline populated, but not congested
- Can happen at many levels
  - Operating system can build up thousands of I/O
  - Can build up at switch ports (buffer credits)
  - Can build up at backend storage ports (inbound queue)
  - Can build up in storage controllers
  - Individual disk devices
    - Native command queuing (NCQ) for SATA AHCI

"Short-Stroking" to reduce latency

- Forcing the use of a smaller area of a rotating disk to reduce seek distance, thus latency
- Less latency means more IOPs
- Penalty is under-utilized storage space

"Advanced Format" (AF) Technology

- Refers to sector size and/or block architecture
  - Expanded from the traditional 512 byte-per-sector and block size formats.
  - Physical disk sector size historically 512/520 bytes
- Limits imposed by various factors, such as MBR structure sizes
- Disk size <previously> constrained to about 2 TB
- "AF" physical sector size is <currently> 4096 bytes (4 kb)
- More space for error checking (CRC)
- More storage space available in same or less physical space
- No corresponding increase in performance capability
- Standards body: IDEMA
  http://www.idema.org/

Advanced Format Continued

- **512 Byte emulation (AF 512e or 512e):**

  Sector size on the media using (e.g. 4,096 bytes-per-sector) while the data passed through the device interface is formatted to 512 Bytes-per-Sector.

- **4K:**

  Shortened form to typically denote 4,096 bytes-per-sector.

- **4K native (AF 4Kn or 4Kn):**

  Sector size using Advanced Format (AF) industry standards to denote that the sector size on the media and the data passed across the host interface from the device interface are formatted to 4,096 bytes-per-sector.

Advanced Format and Partition Alignment

- "Legacy" OS may have storage partitions still misaligned
- The "AF" problem is very similar to RAID misalignment
  - "AF" has error correction fields associated with sectors
  - To modify a physical sector, 3 operations must occur:
    - Same as traditional sector size
      - Read data from disk into buffer
      - Modify (data and error-recovery)
      - Write

Understanding the workload

- Request size
- Burstiness
- "Hot" data
- Concurrency
- Inter-arrival time
  (time of arrival from one request to the next)
- Locality (matters more on rotational than SSD)
- Few customers really understand their workload
- Few tools can faithfully reproduce a given workload

Performance Counters

- Latency
  - Average Disk Sec/Transfer
    - Average Disk Sec/Read
    - Average Disk Sec/Write
- % Idle Time
  - Not always indicative of true idle time
  - RAID or other virtualization can skew results

Frequently Asked Questions

- What are performance considerations associated with Volume Mount Points?
  - Answer: Must use PhysicalDisk counters and must aggregate results. The root mount point does not collect or generate much I/O
- What are the best tools to measure performance
  - Performance sampling tools over time
  - Operating system specific tools
  - Vendor supplied tools
    - Some free, some bundled, some extra cost
  - Analysis services: not free

Refer to Other Tutorials

- Use this icon to refer to other SNIA Tutorials where appropriate

Check out SNIA Tutorial:

OS Storage Performance Analysis

- Use the Hands-On Lab icon only if you've been notified that your tutorial is a cross-referenced match to the Hands-On Lab

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: tracktutorials@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Bruce Worthington PhD