Advanced DataTools Webcast

Webcast on Oct. 20, 2015

Advanced DataTools Webcast Base

Webcast on ??? ??, 2014

International Informix Users Group
www.iiug.org

Informix Storage and RAID5

Doing Storage The Right Way!

Art S. Kagel
ASK Database Management

Quick Review

Really fast tour of chunk storage options vis-a-vis filesystems, etc.

Send a chat to Tom if you want me to go over anything at the end or have any questions during the 'cast!

Options for Chunks

- RAW Device
  - Use character device driver
    - UNIX privs start with 'c'
  - Represents a disk partition, LUN, or other logical storage portion with no OS formatting
  - No OS filesystem overhead
  - OS caching is not available
  - All space used from the device is physically contiguous on disk

Options for Chunks

- COOKED Device
  - Uses block device driver
    - UNIX privs start with 'b'
  - Represents a disk partition, LUN, or other logical storage portion with no OS formatting.
  - OS Caching is normally enabled
  - No OS filesystem overhead
  - All space used from the device is physically contiguous on disk

Options for Chunks

- Filesystem files
  - Uses a normal file entry in a directory
    - First character of UNIX privs is dash '-'
  - Non-contiguous space allocated by the OS within a filesystem mounted from a block (cooked) device
  - OS filesystem overhead applies
  - OS Caching is normally active
  - Filesystem structures must be maintained

Options for Chunks

- SSD – Solid State Disk “Drives”
  - Performance varies widely
  - Drive lifespan and reliability improving all the time
    - Finally in the range needed for serious business use

Options for Chunks

- SSD – Solid State Disk “Drives”
  - Different technologies have different performance and lifespans
    - Flash – Very fast read. Slower writes. Writing decreases life
    - DRAM – Very fast write. Read time not as good as flash
    - Hybrid – Use DRAM as a built-in cache for writing to flash. Use flash directly for reading.

Options for Chunks

- SSD – Solid State Disk “Drives”
  - Different **brands** have different performance and lifespans
    - NAND Flash storage becomes more difficult to read (ie slower) over time due to charge dissipation. Some manufacturers rewrite sectors in the background constantly to refresh the charge levels.
    - Rewrites decrease the lifetime of the NAND Flash memory.
    - Degradation time varies from ~8 to ~40 weeks

Options for Chunks

- SSD – Solid State Disk “Drives”
  - Flash memory performs better at higher temperatures. (Best: ~115°F)
  - Drive controllers perform better at lower temperatures.
  - SSD controllers throttle IO throughput when they begin to heat up negating the advantage of using warmer Flash chips
  - Data Centers are COLD places!

2012-04-23 10:05

Options for Chunks

- SSD – Solid State Disk “Drives”
  - ADTC has tested SSD 'drives' with 11.50 and 11.70
  - ADTC 2010 Fastest DBA Batch Transaction Timing Benchmark
    - Runs 3x faster on SSD on the same machine
  - ADTC 2011 Fastest DBA OLTP Benchmark
    - Runs 3x faster on SSD on the same machine

2012-04-23 10:05

How you do IO matters!

- Prior to v11.10 IDS had to open Cooked Device and File chunks with the o_sync flag enabled to forced synchronous writes in order to insure that data was safely on disk when a write system call returned
- Data was being copied from IDS buffers to OS buffers before being written to disk

How you do IO matters!

- Synchronous writes are slow
- This caused Cooked Device chunks to be ~10-15% slower than RAW Device chunks
- Filesystem chunks were 15-25% slower than RAW

How you do IO matters!

- IDS 11.10 began support for the new o_direct flag which by-passes the OS Cache and writes directly to disk
- Only supported for Filesystem chunks – not Cooked or RAW Device chunks
- Makes Filesystem chunks perform ~5% slower than RAW Device chunks when properly configured

How you do IO matters!

- IDS 11.50 added support for the AIX Concurrent IO model using JFS2 filesystems
- On AIX with o_cio support configured for a JFS2 filesystem and the filesystem properly tuned and mounted, Filesystem chunks are only about 3-5% slower than RAW Device chunks

2012-04-23 10:05

Where you do IO matters!

- Filesystems come in three flavors:
  - “Normal”
  - “Light weight”
  - Journaling

Where you do IO matters!

- General filesystem information -
  - IDS pre-allocates chunks so fragmentation of the initial chunk will depend on the state of the filesystem at the time you create the chunk (except for some journaled FS)
  - Marking a chunk expandable increases the likelihood that the new allocations appended to the chunk will not be contiguous

Where you do IO matters!

- General filesystem information -
  - You have no control over the layout of the chunk's disk allocations within the filesystem
  - Expandable filesystems exacerbate the problem

Where you do IO matters!

- "Normal" filesystems
  - Average overhead, though different filesystems behave differently
  - Most UNIX filesystems attempt to allocate a new file from contiguous disk, but no guarantees for larger files once files have been created and destroyed on the filesystem over time
  - Best to use a brand new filesystem for chunks

Where you do IO matters!

- “Light weight” filesystems
  - Designed by OS and SAN vendors for high speed applications
  - Supposed to be lower overhead than “normal” filesystems
  - Faster file open and close only
  - Anecdotal evidence indicates that they are not noticeably better for database system storage

2012-04-23 10:05

Where you do IO matters!

- Journaling filesystems
  - Adds a form of logging to the filesystem to reduce recovery time and limit data loss if the system crashes with unwritten data in cache

Where you do IO matters!

- Journaling filesystems
  - Many different versions:
    - JFS/JFS2/OpenJFS – IBM developed for AIX and released as open source
    - ZFS – Sun developed for Solaris and released as open source
    - EXT3 & EXT4 – Linux developed
    - BTRFS – Oracle/Linux developed
    - XFS – Silicon Graphics developed for IRIX

Where you do IO matters!

- JFS/JFS2/OpenJFS
  - Journals filesystem meta-data only (not file contents)
  - Serialized writes to maintain data consistency unless Concurrent_IO is enabled
  - Uses variable length extents to build large files
  - Maps/locates extents using a btree index in the inode

Where you do IO matters!

- JFS/JFS2/OpenJFS
  - JFS locks the FS allocation groups during file expansion to improve contiguous allocation. This can block the expansion of other files in that allocation group
  - Fairly low overhead
  - Fairly safe

Where you do IO matters!

- ZFS – Sun developed for Solaris and released as open source
  - Uses copy-on-write transaction model – rewrites are made to a different physical location than data came from and the new block replaces the old one in the meta-data. This causes chunks to become more and more non-contiguous over time

Where you do IO matters!

- ZFS – Sun developed for Solaris and released as open source
  - IDS pages are substantially smaller than ZFS pages. That means many rewrites of the same ZFS page and many relocations that have to be cleaned up in the background

Where you do IO matters!

- ZFS – Sun developed for Solaris and released as open source
  - Dirty block cleanup eats into IO bandwidth and fights applications for head positioning
  - ZFS maintains two to three checksums for every data and meta-data block modifying each up the meta-data tree for every write
  - High overhead
  - Good safety

Where you do IO matters!

- EXT3 & EXT4 – Linux developed
  - Essentially EXT2 with journaling added on.
  - Data and metadata journaled
  - Copy-on-write data rewrites – causes chunks to become increasingly less and less contiguous over time

Where you do IO matters!

- EXT3 & EXT4 – Linux developed
  - EXT4 (& EXT3 with write-back enabled) writes single entry metadata journal entries BEFORE the data block it maps. Can CAUSE corruption if the system crashes before the updated data is written

  Linus Torvalds says: “Whoever came up with (*EXT4's write back policy*) was a moron. No ifs, buts, or maybes about it.”

Where you do IO matters!

- EXT3 & EXT4 – Linux developed
  - EXT4 writes out dirty cache only every 2 minutes (EXT2 & EXT3 do so every 5 seconds)
  - Low safety
  - Low performance

Where you do IO matters!

- BTRFS – B-Tree Filesystem - <NEW>
  - Still under active development, constantly changing with backward compatibility guarantees. (On-disk format stable in kernels after 2014.)
  - Journaled
  - Copy-on-write
  - Indexed directories
  - Snapshot-able
  - Multiple device support - File striping & mirroring
  - Data and MetaData checksums
  - SSD aware and optimized
  - Compression option configurable by file or volume
  - Online expansion/contraction
  - Online balancing of multiple devices by moving blocks in files
  - Built-in RAID levels 0, 1, & 10 (RAID 5 & 6 “experimental”)

Where you do IO matters!

- XFS
  - Meta-data only journaling
  - Write journal before data
  - Dual entry journaling to permit recovery if the modified data is never written
  - Low overhead
  - Good safety

What's behind the scenes matters!

Let's talk about RAID Levels!

RAID – Redundant Arrays of Inexpensive Disks

Proposed in a SIGMOD paper in 1988 because drives were too expensive for mirroring to be practical at the time and drives were not increasing in capacity fast enough to handle data growth.

What's behind the scenes matters!

RAID0 – Striping only
RAID1 – Mirrored drives only

Combinations or RAID0 & RAID1

RAID01 – Mirror two stripe sets
RAID10 – Stripe two or more mirror sets

What's behind the scenes matters!

- NO RAID5! - One extra drive per array to allow for parity blocks. Rotating parity location.
- NO RAID6! - Two extra drives per array to allow for double parity blocks. Rotating parity location
- NO RAIDZ! - Actually ZFS scheme to use copy-on-write to overcome silent disk corruption

2012-04-23 10:05

What's behind the scenes matters!

- NO RAID51! - One parity drive per array and mirror the entire array on another. <LOL!>
- NO RAID61! - Two parity drives per array and mirror the entire array on another. <ROTFL!>

What's behind the scenes matters!

NO RAID5 !!!!!!!!!!!!!!!!
NO RAID6 !!!!!!!!!!!!!!!!
NO RAIDZ !!!!!!!!!!!!!!!!
NO RAID51 !!!!!!!!!!!!!!!!
NO RAID61 !!!!!!!!!!!!!!!!

2012-04-23 10:05

What's behind the scenes matters!

- The following points are now recognized by storage industry studies:
  - Possibility of a second drive failure during recovery is 4x more likely in practice than statistics predict!
    - The safety of RAID5 is predicated on the drive failure rate being low!
  - Server-grade drives have the EXACT SAME failure rates as consumer-grade drives!
    - What are you paying 2-3X the price for anyway?

2012-04-23 10:05

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

- Possibility of a second drive failure during recovery is 4x more likely in practice than statistics predict!

2012-04-23 10:05

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

- The following points are now recognized by storage industry studies:
  - Possibility of a second drive failure during recovery is 4x more likely in practice than statistics predict!
    - The safety of RAID5 is predicated on the drive failure rate being low!
  - Server-grade drives have the EXACT SAME failure rates as consumer-grade drives!
    - What are you paying 2-3X the price for anyway?

2012-04-23 10:05

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

- The following points are now recognized by industry studies:
  - Atomic writes to the multiple drives in an array are not guaranteed!
    - What happens if all of the data and parity drives in a RAID5 array are not all written atomically?

What's behind the scenes matters!

- The following are now recognized by industry studies:
  - Larger drives take longer to rebuild increasing risk of multiple drive failure over conditions in the past when drives were smaller.
  - A recent study by a storage industry association concluded that drives over 1TB are statistically likely to suffer from multiple bit dropouts. Number of bits on the drive exceeds the bit failure rate!

2012-04-23 10:05

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

- The following are now recognized by industry studies:
  - SSD Flash units suffer from 'bit rot' and cosmic ray damage just like mechanical/magnetic disks. More so if they are frequently written to.
  - All in all, the error rates as observed by a CERN study on silent corruption, are far higher than the official rate of one in every 10^16 bits (observed error rates of about one in 10^7 bits ie 1 out of about every 10,000,000 bits or about 1.2MB!)

2012-04-23 10:05

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Correlated failures

- In practice, the drives (in a RAID array) are often the same age (with similar wear) and subject to the same environment. Since many drive failures are due to mechanical issues (which are more likely on older drives), this violates the assumptions of independent, identical rate of failure amongst drives; failures are in fact statistically correlated. In practice, the chances for a second failure before the first has been recovered (causing data loss) are higher than the chances for random failures. In a study of about 100,000 drives, the probability of two drives in the same cluster failing within one hour was four times larger than predicted by the exponential statistical distribution—which characterizes processes in which events occur continuously and independently at a constant average rate. The probability of two failures in the same 10-hour period was twice as large as predicted by an exponential distribution.

http://en.wikipedia.org/wiki/RAID#Problems_with_RAID

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Unrecoverable read errors (URE) during rebuild

- These present as sector read failures. The associated media assessment measure, unrecoverable bit error (UBE) rate, is typically specified at one bit in 10^15 for enterprise-class drives (SCSI, FC or SAS), and one bit in 10^14 for desktop-class drives (IDE/ATA/PATA or SATA). Increasing drive capacities and large RAID 5 instances have led to an increasing inability to successfully rebuild a RAID set after a drive failure and the increasing occurrence of an unrecoverable sector on the remaining drives. When rebuilding, parity-based schemes such as RAID 5 are particularly prone to the effects of UREs as they affect not only the sector where they occur, but also reconstructed blocks using that sector for parity computation. Thus, an URE during a RAID 5 rebuild typically leads to a complete rebuild failure.

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Unrecoverable read errors (URE) during rebuild

- Double-protection parity-based schemes, such as RAID 6, attempt to address this issue by providing redundancy that allows double-drive failures; as a downside, such schemes suffer from elevated write penalty. Schemes that duplicate (mirror) data in a drive-to-drive manner, such as RAID 1 and RAID 10, have a lower risk from UREs than those using parity computation or mirroring between striped sets.

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Increasing rebuild time and failure probability

- Drive capacity has grown at a much faster rate than transfer speed, and error rates have only fallen a little in comparison. Therefore, larger capacity drives may take hours, if not days, to rebuild. The rebuild time is also limited if the entire array is still in operation at reduced capacity. Given an array with only one drive of redundancy (RAIDs 3, 4, and 5), a second failure would cause complete failure of the array. Even though individual drives' mean time between failure (MTBF) have increased over time, this increase has not kept pace with the increased storage capacity of the drives. The time to rebuild the array after a single drive failure, as well as the chance of a second failure during a rebuild, have increased over time.

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Increasing rebuild time and failure probability

- Some commentators have declared that RAID 6 is only a "band aid" in this respect, because it only kicks the problem a little further down the road. However, according to a 2006 NetApp study of Berriman et al., the chance of failure decreases by a factor of about 3,800 (relative to RAID 5) for a proper implementation of RAID 6, even when using commodity drives. Nevertheless, if the currently observed technology trends remain unchanged, in 2019 a RAID 6 array will have the same chance of failure as its RAID 5 counterpart had in 2010.
- Mirroring schemes such as RAID 10 have a bounded recovery time as they require the copy of a single failed drive, compared with parity schemes such as RAID 6, which require the copy of all blocks of the drives in an array set. Triple parity schemes, or triple mirroring, have been suggested as one approach to improve resilience to an additional drive failure during this large rebuild time.

2012-04-23 10:05

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Atomicity: including parity inconsistency due to system crashes

- A system crash or other interruption of a write operation can result in states where the parity is inconsistent with the data due to non-atomicity of the write process, such that the parity cannot be used for recovery in the case of a disk failure (the so-called RAID 5 write hole). The RAID write hole is a known data corruption issue in older and low-end RAIDs, caused by interrupted destaging of writes to disk.

What's behind the scenes matters!

Wikipedia recognizes the following “weaknesses” of RAID:

Write-cache reliability

- A concern about write-cache reliability exists, specifically regarding devices equipped with a write-back cache—a caching system that reports the data as written as soon as it is written to cache, as opposed to the non-volatile medium.

  NOTE: All SAN systems are write-back cache systems!

How big are your disk arrays?

With a URE rate of 1 in 10^14bits or ~12TB if your arrays are bigger than 12TB (whose aren't these days?) second drive failure is statistically PROBABLY not possible PROBABLY!

CERN

The European Council for Nuclear Research

2012-04-23 10:05

What's behind the scenes matters!

- Scientists at CERN beat the hell out of 1.5PB of data on RAID5 and noticed data corruption so they studied it as only physicists can. Their report said:

"The RAID controllers don't check the 'parity' when reading data from RAID 5 file systems. In principle the RAID controller should report problems on the disk level to the OS, but this seems not always to be the case."

2012-04-23 10:05

What's behind the scenes matters!

http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797

Cern tested:

"A <program> was developed <that> writes a ~2 GB file containing special bit patterns and <then> reads the file back and compares the patterns. This program was deployed on more than 3000 nodes...and run every 2 hours. <Five weeks> of running on 3000 nodes revealed 500 errors on 100 nodes."

2012-04-23 10:05

What's behind the scenes matters!

Cern found:

- 80% of their errors were traced to disk firmware bugs
- 10% of errors traced to memory card incompatibility with system boards
- RAID5 could not correct any of these errors nor the remaining 10% due to bit rot (partial media failure and cosmic ray damage – note that Cern is very much underground!).

2012-04-23 10:05

Dell wrote in 2009:

"RAID 5 is no longer recommended for any business critical information on any drive type."

Note: Dell just bought EMC – wonder what they're going to do??

What's behind the scenes matters!

ONLY USE RAID10 (or RAID1). Why?

- Safer from bit rot. Does not copy garbage from drive to drive.
- At least 75% less susceptible to data loss from second drive failure.
- 80% faster recovery time (further reduces 2nd drive failure risk).
- Performance degrades <10% during recovery
    - 6.25% for a 5 pair array versus 80% for a six drive RAID5 array

2012-04-23 10:05

What's behind the scenes matters!

ONLY USE RAID10 (or RAID1). Why?

- Up to 200% higher peak read performance over RAID5. (Cern verified)
- Sustainable 100% increase in write performance over RAID5 without adding huge and expensive cache memory on the SAN.
- Mirroring each drive from a different drive lot reduces the danger from hardware and firmware bugs in the drives. (This was the source of 80% of Cern's data corruption!)

Questions ?!?

2012-04-23 10:05

Doing Storage the Right Way!

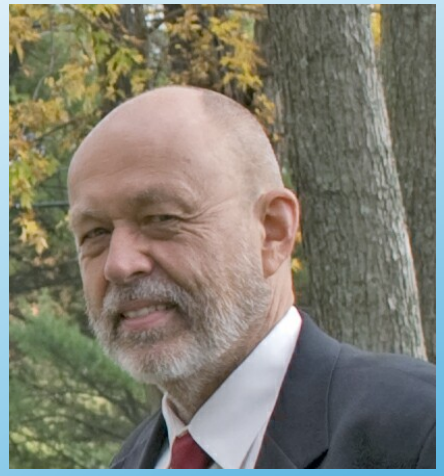
Art S. Kagel

art.kagel@gmail.com
Or
art@askdbmgt.com

2012-04-23 10:05

www.askdbmgt.com

Lester Knutsen

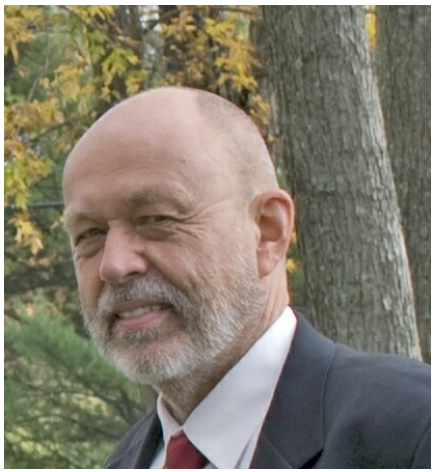


Lester Knutsen is President of Advanced DataTools Corporation, and has been building large Data Warehouse and Business Systems using Informix Database software since 1983. Lester focuses on large database performance tuning, training and consulting. Lester is a member of the IBM Gold Consultant program and was presented with one of the Inaugural IBM Data Champion awards by IBM. Lester was one of the founders of the International Informix Users Group and the Washington Area Informix User Group.

2012-04-23 10:05

lester@advancedatatools.com
www.advancedatatools.com
703-256-0267 x102

Lester Knutsen



Lester Knutsen is President of Advanced DataTools Corporation, and has been building large Data Warehouse and Business Systems using Informix Database software since 1983. Lester focuses on large database performance tuning, training and consulting. Lester is a member of the IBM Gold Consultant program and was presented with one of the Inaugural IBM Data Champion awards by IBM. Lester was one of the founders of the International Informix Users Group and the Washington Area Informix User Group.

lester@advancedatatools.com
www.advancedatatools.com
703-256-0267 x102

International Informix Users Group
www.iiug.org

Mike Walker

Mike Walker has been using Informix databases for 18 years, as a developer and as a database administrator.

Recently Mike has been developing and supporting large data warehouses for the Department of Agriculture.

Contact Info:
mike@advancedatatools.com
www.advancedatatools.com
Office: 303-838-0869
Cell: 303-909-4265

International Informix Users Group
www.iiug.org

Art Kagel

Art Kagel, Principal Consultant of Advanced DataTools Corporation. Art is a member of the IIUG Board of Directors and a recipient of the IIUG Directors Award. Art is a five time recipient of the IBM Information Champion Award. Art has over twenty-eight years of database experience.

He has served as Manager of Database Systems and Services for a major information systems and media corporation and is a leading consultant specializing in IBM's Informix Server.

Contact Info:
art@advancedatatools.com
www.advancedatatools.com
Cell: 732-213-5367

International Informix Users Group
www.iiug.org

Tom Beebe

Tom is a Senior Database Consultant and has been with Advanced DataTools for over 10 years. He has been working with Informix since college with a long time fondness for open source languages. Tom is the lead consultant for Networking, Unix System Administration and Web Development needs. Currently, he is the Project Manager and lead developer on a variety of Web Development projects.

Contact Info:
tom@advancedatatools.com
www.advancedatatools.com
703-256-0267 x 106

International Informix Users Group
www.iiug.org

We are beginning to plan our Webcast for December so keep an eye out for the announcements.

If you have any requests or suggestions for topics you would like to see in future ADTC Webcasts please let us know!

2012-04-23 10:05