Saturday, October 17. 2009

App benchmarks, incorrect conclusions and the Sun Storage F5100

I've just read through several comments regarding the F5100. Is it just me or is common sense on vacation at the moment at some locations. I will go through some of this comments and dissect them here.

Some voices with criticsmA commentator at StorageMojo calling himself "KD Mann". This is my favorite one, because he tries to make a point based on a lot of benchmarks that it doesn't look like the SSDs in the F5100 delivers the performance. When you look at it really cursorily, you may even say "He has a point". But when you really look into the comments and then into the benchmarks it looks different. There are two major problems with this comment: He forgets about an important side. And he can't cite as he swaped results.

I know, that his blog entry gives boths commentators more space than they deserve, but both are good examples how you can fall in a lot of pitfalls while reading benchmarks.

Application benchmarksI should explain something before: Benchmarks are always a mix of task. There are benchmarks with 99% CPU and 1% benchmark related IO (SPECint) and then there are benchmarks with 99% IO and 1% CPU not related to I/O (HDbench for example). You could put dozens of racks of F5100 into a SPECint benchmark configuration without yielding more performance. A benchmark just consisting out of I/O will eat away all your I/O to deliver better results as long the CPUs are capable to move the data around. And then there are a lot of benchmarks in between.

Let's assume a benchmark consists out of 50% IO (reading and writing data) and 50% CPU (doing something with the data). Let's now assume you have a hyper-super-dooper storage reducing the memory access to almost zero. What is the maximum speedup you can expect from using such a device? 50%. With a practical example: You have task that is cpu-bound for 1 minute and i/o-bound for a minute. You need 2 minutes to fulfill the task. Now reduce the I/O to 1 second. You have still a minute and 1 second to go. So what's the maximum speedup? I assume you made the right conclusion.

With multiple threads this the improvement may better than in the example above: As multiple processes contend on the same resource, reducing the average latency keep more processes from waiting on data and doing nothing, while your storage subsystem is searching data for another thread.

More important: Application benchmarks doesn't throw away the data they do computational tasks with it. With this knowledge you should view those benchmarks. The first benchmarks he cites are i/o intensive, but they are even more computational intensive. One of them is finite elements analysis, the other one is MCAE.

Abaqus MCAEFor example take a look at the comments of KD Mann at storagemojo:Got it. An F5100 array of 20 SSDs can outperform an array of six HDDs by 5%.But now let's add some perspective to it: The test cases for the ABAQUS standard module all have a substantial I/O component where 15% to 25% of the total run times are associated with I/O activity (primarily scratch files).
When just 15% to 25% of runtime are related to I/O i think it's pretty nice to accelerate it by 5% the whole runtime due to the usage of SSD. You have to consider here, that those SSD are just capable to reduce the I/O part. You can't do something about the CPU. Let's assume 20% I/O part in average. So this benchmark was 412 seconds in I/O. So the usage of 20 SSD shaved of 92 seconds of this 412 seconds.

NASTRANPretty much the same is valid for the NASTRAN benchmark. It was accelerated by the F5100 by factor 2.1 when compared with a 4 disk RAID 0 configuration. You have to keep in mind that this benchmark is an HPC benchmark, not an storage benchmark. While needing fast storage you have to keep in mind that most of the time this benchmark is CPU-bound and not to I/O. It's mostly an computational benchmark.

So i don't understand why this person uses this benchmark to think that there is something wrong.

PeoplesoftAt last i want to comment the benchmark about Peoplesoft. At first this is a nice example why many benchmark configurations have a large amount of disks: When you look at the benchmark of the Itanium gear he cites, has 58 disks having approximately 8 TB. But they just used 512 GB. Wouldn't wonder if they used a technique called short-stroking. This means you just use the outer sectors of the disks thus reducing the way a head has to move. You can get really nice IOPS values even out of rotating rust when you shorten down the way the head has to go. Would be

nice to see the configuration of the HP EVA to confirm this point.

To the other results: My colleague could have done an error indeed by comparing a "large" to an "extra large" model. But i have not enough data to confirm or to dismiss this. I'm will just waiting for this until the official documentation is available and come back to it.

Nevertheless there are some interesting data points in this benchmark. The sun configuration was able to yield better performance with just 8 instead of 16 threads with 25% instead of 88% CPU-load than the HP solution. You could say, an M3000 would be sufficient to do the job

Capacity is still a factorThe commentator in the BestPerf blog asked why they used 40 FMods while claiming that the F5100 wasn't in sweat. That one is simple. A single FMod has 24 GB. Let's just take the number from the HP disclosure: They stated, that they used 512 GB RAID1 storage: Leads to 1 TB raw capacity. You need 40 FMods for this numbers. We need a little bit less, as we put some stuff to rotating rust disks. It's really that simple. You just have to read the available docs.

Rotating rust for redo logsThe second question of that reader was about the usage of magnetic rust for the redo logs. Well, that's easy too. The F5100 is optimized for writing 4K blocks. The log writer doesn't write in such blocks. And than there is another tooth i have to remove from this reader: Writing the redo log isn't a tough task when it's the only job you do on the disks. Many people forgets in this SSD hype, that rotating rust has nice sequential write/read performance, especially since the introduction of perpendicular recording as more and more bits are moving under the head in a given span of time. It's just the random access that kills those devices. It's similar to the relation hard disks/tape. When a tape drive starts to stream data, you might have problems to feed it with data from your hard disks. It's the winding that kills the performance.

What does the log writer do? Well, just writing log records in an arbitrary size. You never read them, okay ... almost never: You read them when your database has gone down in flames. So it can be quite reasonable to use magnetic rust for a dedicated log writer filesystem and SSD for your datafiles, as the access patterns to that files are much more random and head movement and rotational latency become a larger factor. If you don't believe me, just dtrace your log writer. BTW: There is a great articles of my colleague Volker Wetter in regard of I/O on Oracle in the Sun Wiki. You should look at "Getting insights with DTrace - Part 1". Given the fact that SSD may be optimized for certain block sizes, it may be a reasonable choice to use rotating rust especially when the flash device is optimized for other tasks. In regard of the nature of performance impacts to LGWR speed you should read Kevin Clossons "Manly Men Only Use Solid State Disk For Redo Logging. LGWR I/O is Simple, But Not LGWR Processing", which give some interesting insights into this topic.

There are several good use cases for SSDs like the flash-extended SGAs in the newes versions of Oracle or the indices. But redo logs aren't the best choice in 99%.

ConclusionAt the end even small gains of performance are the expression of large increases on one area. In any case you should really look deeply into a benchmark before making statements about something.

```
Posted by Joerg Moellenkamp in English, Oracle at 13:11
```

First off, KD Mann was analyzing Sun's own benchmark cites, not like he pulled them out of thin air.

Secondly, Mann didn't swap the cites, Sun did.

Go look at the numbers he cited and compare to Sun's posts...
Anonymous on Oct 18 2009, 18:03

Yes, but he drew incorrect conclusions out of it.
Anonymous on Oct 18 2009, 20:24

Re: "58 disks having approximately 8 TB. But they just used 512 GB."

Why does it matter if the test only requires a 512GB database...the whole question is based on IOPS...why did the Sun/Flash system need 40 SSDs and 12 HDDs (all RAID-0), which costs five times as much as the HDD setup, in order to deliver merely the same performance as the 58 HDDs?

FYI, 40 x 24GB SSDs plus 12x450GB HDDs is also many times more capacity than the 512GB required for the benchmark. Shouldn't the random I/O have been easily handled by only six SSDs? Or does Sun also need to short-stroke it's SSDs (for decent write performance) the same way Intel does?
Anonymous on Oct 20 2009, 18:30

Actually, Sun themselves more or less validates these observations...read through the comments for Sun's replies...

http://blogs.sun.com/BestPerf/entry/oracle_peoplesoft_payroll_sun_sparc

Regarding Flash for log files...Intel research reports that a single HDD with on-disk cache enabled is 2x faster than Intel's own X-25 SSD for log files. It looks like this is because Flash write performance sucks just as badly as Sun's Michael Cornwell has said.

http://portal.acm.org/citation.cfm?id=1559855
Anonymous on Oct 20 2009, 20:05

Jorge, I think you should slow down here.

In an apparent attempt to defend Sun's honor, you've glossed over the essential point these folks are trying to make, and it's a good one. What is the business- case for Flash SSD, and has it been quantified?

From what I can see -- it's a very good question.

Also, you should note that the excellent post by Kevin Closson you cited (a) doesn't really support your point, and (b) referred to testing using the Texas Memory Systems DRAM based SSD, not Flash. DRAM SSD and Flash SSD are two completely different animals

Sure enough, the Intel paper above indicates that a single HDD with WCE (they called it the "ideal" case) is twice as fast as the X25 Flash based SSD, though I had to go to the footnotes to find out which SSD the Intel guy was referring to...
Anonymous on Oct 20 2009, 20:37

Can anyone here direct me to the published Peoplesoft Payroll result with the Sun F5100? Oracle has the HP-based 58xHDD system posted on it's Peoplesoft site, but the Sun result is nowhere to be found.

I'm always suspicious when companies cite and blog about benchmark results that they haven't yet published -- usually there is something they want to hide.
Anonymous on Oct 20 2009, 20:44

1. The problem is: You can't use a disk with enabled write cache for data base logfiles, as this would break the ACID of databases. Lose you power, lose the cache, lose transactions.

2. The X-25E is slow drive at writes, especially when you have to obey write barriers as the X25E has no cache protection.

3. The observation that disks can be really fast is in the nature of writing logs. It consists out of sequential writes. And disks are really fast a writing things sequential. They just suck at random reads.
Anonymous on Oct 21 2009, 10:12

It takes some time until the white paper is out. But i wouldn't hold my breath about this: When Sun doesn't publish the benchmark paper, this will will be an PR disaster. And when Sun publishes them, and the data is proofed as wrong, it will be an PR disaster as well. So you can cool down on that ...
Anonymous on Oct 21 2009, 10:16

Steve A posts, "Why does it matter if the test only requires a 512GB database...the whole question is based on IOPS..."

The short-stoking by a factor of 16x is deceptive.

As the database expands to consume more space, the performance of the system will severely degrade. In this case.

The power, cooling, and rack space required for the solution is 16x the real capacity required for the solution. If the business needs to double the size of the database, the power/cooling/space requirements would increase by another factor of 16 over the real capacity requirements.

"why did the Sun/Flash system need 40 SSDs and 12 HDDs (all RAID-0), which costs five times as much as the HDD setup, in order to deliver merely the same performance as the 58 HDDs?"

To save massive quantities of rack space, hvac costs, and power costs.

"FYI... is also many times more capacity than the 512GB required for the benchmark."

Achieving similar performance using an order of magnitude fewer racks of equipment is _not a bad thing_.

It would be nice to see a performance chart, of some kind, as SSD's are scaling up, with some kind of rule-of-thumb to measure what percentage of SSD's are required for a quantity of data storage as well as IOP's percentage from optimum... so people architects could optimize architecture for a particular budget.
Anonymous on Oct 21 2009, 19:33

Ickabod asks, "What is the business-case for Flash SSD, and has it been quantified?"

I suspect the business case has to do with reoccurring costs related to: applications/database license & maintenance savings, hvac, power, land tax (space savings), and future CO2 tax savings by the global-warming folks.

I am interested in seeing some analysis on this!

Anonymous on Oct 21 2009, 19:46

Ickabod asks, "What is the business-case for Flash SSD, and has it been quantified?"

David comments, "I am interested in seeing some analysis on this!"

I found something on this.
http://mediacast.sun.com/users/ClaudiaHildebrandt/media/ZFS_arc.pdf
Page 36
- 4% more cost
- 3.2x better read iops
- 11% better write iops
- beats storage wattage by 4.9x
- 2x greater raw capacity

I think 4% increase in cost for double the storage capacity is good enough economics by traditional metrics.

The additional performance increase is just icing on the cake!

OK - sees the business case question has been solved.

I would still like to see some nice graphs showing where the sweet spots are for cost-benefit analysis!

Anonymous on Oct 21 2009, 21:13

David,

I appreciate the link, but it's not at all helpful. (anyone reading along here can go to slide 36 of the presentation at http://mediacast.sun.com/users/ClaudiaHildebrandt/media/ZFS_arc.pdf to see for yourselves).

1) These are just IOmeter synthetic benchmark numbers, not application performance results. We already know that SSD's do 10x to 100x better on IOmeter than they do in applications like TPC-C or Peoplesoft Payroll. I am looking for real-world application benchmarks like the ones Joerg is commenting on here.

2) the "power savings" numbers are ridiculous and deceptively contrived...you do not get 80% system power savings by replacing 7x10K HDDs (@9w each) with 5x4200RPM HDDs plus 2xSSDs!!! System power savings here are less than 25 watts and the system uses at least 600 watts -- that's about 4.1% improvement, not 4,900%

3) If I needed more capacity, I could switch out the 146GB disks for 300GB disks for a total cost increase of about $250.

This is just more BenchMarketing malarky.

In Oracle's own Peoplesoft Payroll benchmark, the SSD system costs at least 5x more and delivers no better performance than the HDD system. Energy savings are miniscule compared to capex required -- and the system will take 20 years or more to pay for itself in energy costs. I am still looking for someone to point to any cost/benefit for SSD in actual, realistic user deployment scenarios.

Anonymous on Oct 21 2009, 23:54

David...the HP based 58xHDD system uses 2.5" disks and fits in a total space of 11RU. The Sun SSD-based system requires 7RU.

How on earth do you concoct a requirement for 16X more rack/power/space vs. the SSD system???

The Sun SSD platform costs 5X more. It doesn't go faster. It requires me to manage multiple "performance islands" of storage. It doesn't use much less power.

And did I mention is costs 5x more for no performance improvement?

Finally, have you ever tested the performance gains from short-stroking? If you had, you'd know that you only get 25% more IOPS maximum by using a 5% stroke vs. a typical 2/3rds stroke. This 512GB database could increase to 4TB and you'd never notice the performance difference.

David...I'm guessing you've never architected or deployed actual application platforms, nor worked with IT decision-makers, have you?

Anonymous on Oct 22 2009, 00:15

Joerg, I was surprised myself, but Oracle uses HDD with write-cache enabled for log files already, they just mirror them -- which also mirrors the cache...and meets TPC-C ACID durability requirements.

FYI, Oracle did this with HP to get the world's #1 lowest cost/transaction TPC system (1/5th of the F5100 cost/transaction) here:

http://www.tpc.org/tpcc/results/tpcc_price_perf_results.asp

See page 16 and bottom of page 19 here for the log-disk cache disclosure.

http://www.tpc.org/results/FDR/TPCC/HPML350G6OELTPCC_FDR.pdf

Anonymous on Oct 22 2009, 00:31

Honestly, reading all of this, I think Oracle is going to be the death of Sun, not a savior.

Sun has been a truly amazing force for innovation over the years, THE thought leader in computing. All of this Oracle benchmarketing garbage and noise threatens to make Sun out to be nothing more than purveyor of SSD nonsense.

Leaves me nostalgic for the days where Sun and MySQL were the sand in Oracle's vaseline.

sigh....
Anonymous on Oct 22 2009, 00:42

To the persons behind the entities "Steve A." and "Ickabod": Normally i wouldn't disclose any data about the commentators. But in some cases there are some irregularities that are too obvious and i think, almost two thousand subscribers have the right to know about it when there is something potentially fishy. To protect your privacy i will obfuscate the information, but the i have unmodified logfiles and mails.

Before I'm going to dig into the point i should explain that S9Y writes the IP of a commentator into a mail informing me of a new comment. This IP number is at the same position every time. The normal pattern is a change in this line, so a line staying equal is an abnormal pattern and a trained eye recognizes this immediately. And then I'm getting really curious!

So:

1. Would you like to explain, why the entities "Ickabod" and "Steve A." have the same IP address in a Dial UP IP-Segment as specified by Spamhaus?

2. Would you like to explain, why the entities "Ickabod" and "Steve A". are using a really long string identifying the browser that is absolutely equal?

aa.bbb.ccc.ddd - - [22/Oct/2009:00:31:15 +0200] "POST /index.php?url=archives/6021-App-benchmarks,-incorrect-conclusions-and-the-Sun-Storage-F5100.html HTTP/1.1" 302 20 "http://www.c0t0d0s0.org/index.php?url=archives/6021-App-benchmarks,-incorrect-conclusions-and-the-Sun-Storage-F5100.html" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; Tablet PC 1.7; .NET CLR 1.0.3705; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

"aa.bbb.ccc.ddd - - [22/Oct/2009:00:42:25 +0200] "POST /index.php?url=archives/6021-App-benchmarks,-incorrect-conclusions-and-the-Sun-Storage-F5100.html HTTP/1.1" 302 20 "http://www.c0t0d0s0.org/index.php?url=archives/6021-App-benchmarks,-incorrect-conclusions-and-the-Sun-Storage-F5100.html" "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; Tablet PC 1.7; .NET CLR 1.0.3705; .NET CLR 1.1.4322; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)"

Any good explanation would be really appreciated!
Anonymous on Oct 22 2009, 09:30

Joerg,

I use a screen name when it suits the situation. In this case it was to entice you to respond substantively rather than dismissively. You began to respond substantively only after Ickabod chimed in.

You said..."two thousand subscribers have the right to know about it when there is something potentially fishy."

Ok...Sun's website says a single FMOD SSD is as fast as 100 disks, yet Sun and Oracle cannot point to a single application where this really happens. Something smells fishy.

That smell get's stronger here, where Sun says the 5X more expensive F5100 solution delivers 14% faster results on PeopleSoft Payroll:

http://blogs.sun.com/BestPerf/entry/why_sun_storage_f5100_is

And here, where they now claim 33% faster on the same benchmark:

http://blogs.sun.com/BestPerf/entry/oracle_peoplesoft_payroll_sun_sparc

When Sun's bloggers are informed (and acknowledge) that the reality is only 2% faster...nobody at Sun bothers to correct the numbers.

Perhaps you will agree that hundreds of thousands of IT customers have the right to know when Oracle and Sun are potentially practicing organized deception -- isn't that also fishy?
Anonymous on Oct 22 2009, 18:27

Disgusting. Simply disgusting. You fake a discussion, just to give other the impression that others share your point of view.

I'm preparing a comment to respond to the partly just outright nonsensical comments you are placing in the comments. But because of all the bs, this takes a while.
Anonymous on Oct 22 2009, 18:56

Joerg, here's one thing we agree on. You said:

"In any case you should really look deeply into a benchmark before making statements about something."

From Sun's F5100 performance page:

==========

"Disclosure Statement
Oracle PeopleSoft Payroll (NA) 9.0 benchmark, Sun M4000 (4 2.53GHz SPARC64) 79.35 min... HP rx6600 (4 1.6GHz Itanium2) 105.70 min, www.oracle.com/apps_benchmark/html/white-papers-peoplesoft.html Results 10/13/2009."

==========

http://www.oracle.com/apps_benchmark/doc/peoplesoft/performance-report/ps9-na-pay-9_ora_hp_rx6600.pdf

Funny...I just noticed that the HP system was not 105.70 minutes...it was 68 minutes. That's eleven minutes faster than the F5100 result.

I guess Sun should "really look deeply into a benchmark before making statements", or at least look at the first page...that's where the 68 minutes number is found.

Joerg, I'll leave the topic now, and let you have the last word -- it's your blog after all.
Anonymous on Oct 22 2009, 19:12

This will be a rather long comment, it will be a mass answer to several comments to two entities of a person, who did it "to get my attention". Yes, suuuure. I think, it was to fake a discussion to make the impression, that people where sharing his point. Don't ask me about his incentive to do so.

This comment isn't for this reader, it's for all the readers stumbling over his comments. It would be easier simply to delete them, but that would give his statements vastly more credibility than they deserve.

I would like to answer to his comment #6.1.1.1 at first. The 33% statement was computed against a benchmark resulting in a 105 minutes runtime on the HP side and 79.35 for Sun. The document isn't available at the Oracle site any longer. But there is a new one. It dates to the Fri, 09 Oct 2009 18:48:28 GMT document as stated by the HTTP header and by Acrobat Reader when you ask it for the document properties. The statement in the blog with the 33% was made on 13 October. I just assume that the page was updated shortly after the blog comment. I just assume that the the document has to go through the same approval processes at HP and then at Oracle thus this sounds reasonable that it took a few days to get it on the webpage. At the moment i just can say: HP, that was really well timed. But: I never thought that HP are idiots. But there is nothing fishy about that.

The second point he tries to make there is that HP is indeed faster with mentioning this 67 minutes number. But that is the wrong number. You have to add in the 13 minutes for "Print Advice" and "Deposit". Then you have the numbers you can compare.

I don't have an idea what they did to accelerate the same system with roughly the same storage. My educated guess the secret lies in the increase of threads to 16. The Sun solution used 8 threads. I would really appreciate if someone sends me the document that was on the website before Friday 9th.

Let's go further in the thread: In several statements Steve A. tries to make the impression that the HDD solution is much cheaper. Well, when you read into the documents he cites even himself, you will see, that the storage system used by HP isn't just a mere JBOD. It's an EVA8100. We don't talk about an 58 HDD, we talk about 58 disks in a modular storage system on the higher end of HP storage product portfolio. Bigger than a 8100 is just their XP series and that's rebranded Hitachi High End Storage. For example with 8 GB cache in a pair of controllers, of course backuped by batteries so they can ignore CACHE FLUSH commands. In addition there is still my opinion, that this configuration is severly short stoked. But to make this short: This device is expensive. It's worth it's money. I don't have a price at hand for the 8100 but the current model, the 8400 starts at $61,456 no disks included list price AFAIK. Then shop for 58 146 SAS disks plus 5 trays. I don't have an idea why Steve A. thinks that the SSD solution is 5 times more expensive than the HDD solution.

So ... let's get to comment #3.1.1. In this article, Steve A. found it nescessary to insult David. But it starting with a misunderstanding. Steve A. talked about the Peoplesoft benchmark whereas David talked seemingly about the TPC-C benchmark. But that's not the point. But the point made by Steve A. has an errornous foundation. He talks about 11U needed by the HP solution and that this solution used 2,5" drives. Well ... i don't know what HP EVA8100 he uses, but in my QuickSpec sheet the 8100 is specified with 14 disks per tray and a tray height of 3 RU. I just assume he got confused by the 146 GB per disk numbers. There were such disks in 3,5" a while ago For 58 disks you need 5 trays. 15 rack units without controller. Add 4 Rack units for the controller, now we are at 19 RU. At the moment it's 3 RU (1RU F5100 + 2RU J4200) versus 19 RU. In my view of the world that's pretty significant. Additionally: Given that a 2C6D EVA8100 consumes 2600 Watts and a 2C2D at 1150 Watts, it would a benevolent assumption to consider a 2C5D with a almost unpopulated fifth shelf in the range of 2000 Watts. Now let's turn to the Sun side:The F5100 is rated at 281 Watts when used at 100% load, 220 Watts at 50% load. Let's just assume 250 Watts for the load at this benchmark (albeit i would assume the power consumption is near the idle load), the J4200 with 15k SAS with two SIMs at 352 watts. 600 Watts versus 2000 Watts. In my world this is pretty significant, too. 1400 Watts. Overthethumb-calculation: 73584 kWh in three years (assuming the rule 1 Watt A/C for 1 Watt into the storage)

When we step to comment #3, Steve A. shows a clear sign of "I can read, but i can't understand". At first i should say, that SSD isn't a silver bullet and there are situations where an HDD is as fast as a SSD. One of this cases is the sequential write. Due to developments like perpendicular recording a hard disks can consume a lot of data. There is only one thing that kills a hard disk: Moving it's head. The configuration used by Sun mirrors this. The indices, the data files have a random access pattern, thus it's best to use SSD. The redo log is steady steam of sequentially written data in arbitrary sizes, a rotating rust disk is perfectly for this task, as those disks doesn't see any random I/O.

To answer his comment: It's about capacity, and even a blind man should see that. When the dataset is 200 GB large, you have to store it somewhere. The hard disks aren't used for the datafiles, just for the redo log, thus you have to keep them out of the capacity calcluation and thus you need the 40 FMod instead of the 20 FMod configuration of the F5100. Simple math. Steve A. doesn't seem to get this, albeit my colleague Vince is stating this clearly in http://blogs.sun.com/BestPerf/entry/why_sun_storage_f5100_is

The next example of "I can read, but i can't understand" is the comment #4: Of course a RamSAN and the F5100 are two pretty different implementations of a SSD. But that wasn't the point of his article. The point was as far as i understood it: It isn't effective to throw SSD at logging. There are different problems with the LGWR limiting its speed. And that pretty much supports my point, that there is no point in using SSD for log writing and that HDD are up to this task. Sun uses this hybrid approach at many occasions. For

example the HSP of ZFS is the same. SSD where it fits (L2ARC,sZIL), HDD where it fits (Pool)

The problems with the HP TPC-C benchmark (It's a HP benchmark, not a Oracle one) as stated in #2.1.2.1 is not really understandable to me. As long a system complies to the ACID rules by the TPC-C it's okay. The large caches on the controller are battery-protected. By the way ... there is a reason, why many TPC-C benchmarks contain a lot of UPSes in their bill of material. It has something to do with write-caching to a large part;) But i'm not here to defend a HP benchmark result ....

Regarding this comment at #4.1.1.1: Hmm ... how do do i say this politely. It was a little bit unfortunate to cite Claudias presentation, it was more a presentation about ARC, L2ARC and sZIL, but spiced up with some marketing-slides. I use them since last year (when my memory serves me right). It's a little bit outdated technology wise, but the core of should be still true, when you update it with current technology: Yes, there are bigger high-RPM disks, but there are bigger low-RPM disks as well and so on. And the power calculation didn't included the server, as this one is equal in both configurations. But that's point where i have to admit that Steve A. are right: This slide needs a recalculation.

So ... it's late now ... i spend my evening and a large heap of tea on writing this comments. There are many more points i would like to comment, but time isn't infinite and i think i already spend more time on it than this person with such a disgusting style in discussion deserves.
Anonymous on Oct 22 2009, 22:54

Hi Ickabod / Steve A.

> I appreciate the link, but it's not at all helpful...

It was helpful to a great deal of people, to understand the performance characteristics of flash in conjunction with using lower cost & higher capacity disks.

> I could switch out the 146GB disks for 300GB...

Yes, and you could switch out the large capacity slower drives for 2TB drives in the flash scenario. The scenario illustrated was helpful in understanding the benefit of flash.

> In Oracle's own Peoplesoft Payroll benchmark...

If you are going to make a weird claim, please provide a link to substantiate it, so there is something objective to truthfully discuss.

> This is just more BenchMarketing malarky.

If you are interested in something besides malarkey - spell correctly, identify yourself consistently, cite objective comparative work, and be truthful with people.
Anonymous on Oct 24 2009, 17:31

David, regarding your most erudite and constructive comment:

>>"...if you are intersted in something besides malarkey - spell correctly.

Huh?

1) http://dictionary.reference.com/browse/malarky

2) http://www.merriam-webster.com/dictionary/malarky

3) http://www.thefreedictionary.com/malarky

There...now you are full of references to both of the correct spellings of malarky/malarkey.

Then you said: "If you are going to make a weird claim, please provide a link to substantiate it, so there is something objective to truthfully discuss."

You'll find it referenced above...it's that long stringy-thing I posted that begins with "http://www.oracle.com/apps_benchmark..." and ends with "...performance-report/ps9-na-pay-9_ora_hp_rx6600.pdf"

Now...do you care to actually address the points I raised about the Sun marketing fluff you posted?
Anonymous on Oct 24 2009, 20:17

I think your point has been addressed by #6.1.1.1.1 ....
Anonymous on Oct 24 2009, 21:06

Hi Ickabod / Steve A.

> the HP based 58xHDD system uses...

I am not sure who cares. Please cite an objective case study comparing them for discussion.

> How on earth do you concoct a requirement for 16X more rack/power/space vs. the SSD system???

You referenced the TPC council as your home page. The latest TPC-C benchmarks cited by Sun & Oracle demonstrate this.

> The Sun SSD platform costs 5X more. It doesn't go faster. It requires me to manage multiple "performance islands" of storage. It

doesn't use much less power.

According to the latest TPC-C benchmark, you are incorrect, yet you cite the TPC as your homepage.

> And did I mention is costs 5x more for no performance improvement?

According to the TPC-C benchmark, you are incorrect.

> Finally, have you ever tested the performance gains from short-stroking?

The published TPC-C benchmark showed performance differences between short-stroking and flash. One does not eliminate the other, flash offers a lower life cycle cost alternative.

> I'm guessing you've never architected or deployed actual...

You guess wrong. Then again, you don't even know your own name. Figure out who you are before you guess wrongly about someone else.
Anonymous on Oct 26 2009, 19:33

David,

When I said...

>> the HP based 58xHDD system uses...

...then you said:

>I am not sure who cares. Please cite an objective case study comparing them for discussion.

Have you read this page?

The "objective case study comparing them for discussion" you ask for was presented by Sunacle themselves! It's the topic of Joerg's entire bloody post here!!

Here it is again, exactly like it was pasted above.

http://blogs.sun.com/BestPerf/entry/oracle_peoplesoft_payroll_sun_sparc

If you had been reading along, you'd know that the entire discussion on this page is about Apps benchmarks and Sun's (incorrect) boast that F5100 based system was faster than the HP 58HDD system on Oracle's PeopleSoft benchmark, when in reality it was slower.

From Sun's F5100 performance page:

==========

"Disclosure Statement
Oracle PeopleSoft Payroll (NA) 9.0 benchmark, Sun M4000 (4 2.53GHz SPARC64) 79.35 min... HP rx6600 (4 1.6GHz Itanium2) 105.70 min, www.oracle.com/apps_benchmark/html/white-papers-peoplesoft.html Results 10/13/2009."

==========

http://www.oracle.com/apps_benchmark/doc/peoplesoft/performance-report/ps9-na-pay-9_ora_hp_rx6600.pdf

Repeating for your benefit...the HP/HDD system was not slower than Sun/SSD, it was eleven minutes faster.

You said:

>>The latest TPC-C benchmarks cited by Sun & Oracle demonstrate this.

No, they don't demonstrate anything here. Comparing a TPC-C result with a PeopleSoft payroll benchmark is like comparing chalk and cheese. Payroll is a "batch mode" application and TPC-C is an online transactional application. One is a synthetic benchmark and the other is a real-world application benchmark.

Perhaps you did not notice that the title of this post is "App benchmarks..."

So...are you (David) ever going to address the question: Why does it take 40 Sun SSDs PLUS 12x15K HDDs to equal the real-world, application performance of a system that just uses 58xHDDs, no SSDs, and costs a fraction as much?

Getting a substantive response here is like trying to nail jello to the wall.
Anonymous on Oct 27 2009, 23:05

Hey, are you unable to read when someone is answering to you. I will cite my comment #6.1.1.1.1 : "So ... let's get to comment #3.1.1. In this article, Steve A. found it nescessary to insult David. But it starting with a misunderstanding. Steve A. talked about the Peoplesoft benchmark whereas David talked seemingly about the TPC-C benchmark. But that's not the point. But the point made by Steve A. has an errornous foundation. He talks about 11U needed by the HP solution and that this solution used 2,5" drives. Well ... i don't know what HP EVA8100 he uses, but in my QuickSpec sheet the 8100 is specified with 14 disks per tray and a tray height of 3 RU. I just assume he got confused by the 146 GB per disk numbers. There were such disks in 3,5" a while ago For 58 disks you need 5 trays. 15 rack units without controller. Add 4 Rack units for the controller, now we are at 19 RU. At the moment it's 3 RU (1RU F5100 + 2RU J4200) versus 19 RU. In my view of the world that's pretty significant. Additionally: Given that a 2C6D EVA8100

consumes 2600 Watts and a 2C2D at 1150 Watts, it would a benevolent assumption to consider a 2C5D with a almost unpopulated fifth shelf in the range of 2000 Watts. Now let's turn to the Sun side:The F5100 is rated at 281 Watts when used at 100% load, 220 Watts at 50% load. Let's just assume 250 Watts for the load at this benchmark (albeit i would assume the power consumption is near the idle load), the J4200 with 15k SAS with two SIMs at 352 watts. 600 Watts versus 2000 Watts. In my world this is pretty significant, too. 1400 Watts. Overthethumb-calculation: 73584 kWh in three years (assuming the rule 1 Watt A/C for 1 Watt into the storage)"

Your 58 disks are an EVA8100 ... and this device that is surely not cheaper, takes an awful lot more of power and space.

The RAID controller alone needs more space than the whole storage at the Sun solution ...
Anonymous on Oct 27 2009, 23:47

Joerg/David,

I am talking about the underlying fundamentals of the storage technologies used here. these are:

- HDD based system:
58x 15KRPM disks

- SSD based system = 40xSSD + 12x 15KRPM disks

Based on the fundamentals, the SSD based system costs several times more and SSD was eleven minutes slower.

For this price calculation, I am using Sun's per-disk pricing for their best HDD system and comparing it to the F5100/FMOD pricing.

http://www.storageperformance.org/results/a00068_Sun_J4400_executive-summary.pdf

So...I can connect 58x 15KHDDs using J4400/J4200 for an all-included cost of $29,000-- including 3yrs of Sun Gold Service Maintainence.

And it fits in 7U.

So...are you (David (or Joerg)) ever going to address the question: Why does it take more than 40 Sun SSDs PLUS 12x15K HDDs to equal the real-world, application performance of a system that just uses 58xHDDs, no SSDs, and costs a fraction as much?

Getting a substantive response here is like trying to nail (warm) jello to the wall.
Anonymous on Oct 29 2009, 18:54

1. You get really annoying. Are you that way by purpose or by nature ?
2. You may put 58 disks in 7U, they may be cheaper, but the benchmark wasn't done with such a device. Read the benchmark. That's the simple error in your thoughts
3. The benchmark you mentioned was done with a EVA8100.
3.1. Do you know the meaning of this little fact you omit in every comment ?
3.2. Do you know what a HSV210-B is ? There are two of it in a EVA8100.
3.2. Do you know for example of the vast amount of cache in DRAM in front of the harddisks?
3.3. Do you really believe that you get the same performance from a JBOD than from Enterprise Storage?
4. Did you really not understand, that this benchmark needed 40 FMods for Capacity?
Anonymous on Oct 29 2009, 19:23

Joerg,

1. Your logical fallacy is covered here ...http://www.fallacyfiles.org/loadques.html Your question implies your answer. Logical fallacies like this are generally used to distract the audience when the debate isn't going well.

2. Sun is selling SSD FMODS at $1700 for 24GB (landed cost in F5100 JBOF). The profit margins are ridiculous and they don't deliver better cost/performance. This is about the SSD ripoff that all storage vendors are perpetrating today.

3. So what? The Sun J4400 delivers more SPC-1 IOPS-per-HDD than EVA, and has among the world's lowest cost/IOPS.

3.1. I "omit this little fact" because it is not only "little", it is utterly meaningless in a debate about the fundamental merits of HDD vs. SSD. FYI, 15K HDDs in an EVA actually perform only half as good on an IOPS-per-spindle basis (110-140/spindle) than they do in a setup like the Sun SPC-1 HDD result I referred you to (300/spindle).

http://www.wmarow.com/storage/spc1.html

Potentially this means that if Sun customers stick with Sun HDD storage, they will get this same peoplesoft performance with far less than 58 HDDs, and at 1/5th the cost of Sun SSD.

3.2.2 Did YOU not know that F5100 used 64MBx40=2.5GBytes of DRAM write cache in front of the SSDs? That's more than EVA uses in it's SPC-1 tests.

http://www.sun.com/storage/disk_systems/sss/flash_modules/specs.xml

3.3. re: "Do you really believe...". No, I don't believe, I know...that JBOD, tightly coupled to the application always outperforms "Enterprise Array" controllers. In case you hadn't noticed, that's why you have never seen "Enterprise Storage" arrays used in any competitive TPC-C result...ever. Sun's own benchmark results prove it. Realworld applications benchmarks like the ones we're talking about here prove it as well. And...what is "JBOD" except for a meaningless distinction/distraction about where the RAID or striping controller or logic goes? And isn't the F5100 "JBOF"?

4. You keep repeating "...this benchmark needed 40 FMods for Capacity" No Joerg, you are grasping at straws and ignoring the facts;

(a) the size of the entire database for Peoplesoft 9.0 benchmark is only 200GBytes (as you would see if you had "carefully examined" the HP result), so even if Sun mirrored (??) the 40x24GB SSDs, they still had more than 2x the necessary storage. (b), the 12x450GBx15KRPM HDDs that Sun used (in a J4200 JBOD, FYI) provided at least another 2.7TBytes.

Joerg, I have a suggestion. Why don't you call or e-mail Jignesh Shah, because he's the guy who is actually doing Sun's Peoplesoft Payroll benchmarks, and he has the actual test results that Sun has not published. Ask Jignesh to release them to be posted and published on the Oracle Peoplesoft site (like the HP result is), and then we can sort this all out, right?

Jignesh is at J.K.Shah@Sun.COM
Anonymous on Oct 30 2009, 16:25

"that JBOD, tightly coupled to the application always outperforms "Enterprise Array" controllers. In case you hadn't noticed, that's why you have never seen "Enterprise Storage" arrays used in any competitive TPC-C result...ever. Sun's own benchmark results prove it."

Actually, the historical reason TPC-C configurations have largely not included Storage Arrays is due to cost. More results are coming out these days with arrays as opposed to piles of JBOD. Consider IBM's 6 million+ TPC-C from the summer of 2008. This used 68 TotalStorage DS4800 (Engenio rebadged if I recall correctly):

http://tpc.org/results/individual_results/IBM/IBM_595_20080610_ES.pdf
Anonymous on Nov 6 2009, 22:24

Kevin,

Aren't Oracle's ExaData machines (both 1 and 2) examples of "JBOD tightly coupled to the application?"

Isn't the Sun/Oracle TPC-C an example of JBOF(lash) in "direct-attach" mode?

If you look at the SPC-1 results, you'll find that (for example) the J4400 Sun JBOD in server-direct-attach mode extracts more SPC-1 IOPS-per-spindle than just about anything else ever tested.
Anonymous on Nov 8 2009, 19:08

Kevin, re: "More results are coming out these days with arrays as opposed to piles of JBOD."

That's not the case.

Let's take a look at the most recently published TPC-C results.

http://www.tpc.org/tpcc/results/tpcc_last_ten_results.asp

Of the five most recent results (all published in 2009), all of them used JBOD attached to the server(s) -- no enterprise arrays.

These systems (on average) extracted about 1,100 TPMc per HDD spindle (or 1,600 per SSD in the case of the F5100 Sun/Oracle JBOF result).

Of the ten most recent results, the only ones (3) that used "enterprise arrays" were from 2008 and delivered (on average) only about 600 TPMc per spindle.

That means that for a given array size (spindlecount), the JBOD based systems will be almost --TWICE-- as fast.
Anonymous on Nov 10 2009, 15:44

Steve A. / Ickabod, "In case you hadn't noticed, that's why you have never seen 'Enterprise Storage' arrays used in any competitive TPC-C result...ever"

Kevin, "More results are coming out these days with arrays as opposed to piles of JBOD."

Steve A. / Ickabod, "That's not the case... Of the ten most recent results, the only ones (3) that used 'enterprise arrays' were from 2008..."

Steve A. unwittingly demonstrates Kevin's point --- 3 out of 10 are "more results" than 0 out of 10... and conveniently forgot Sun scored a #1 position in TPC-C using their arrays.
http://www.oracle.com/us/corporate/press/036544

Steve A. / Ickabod, "for a given array size (spindlecount), the JBOD based systems will be almost --TWICE-- as fast"

Your conclusion based upon taking unrelated data points is tremendously flawed.

Similar reasoning would make disk drive manufacturers reduce the cache on their disks and would make consumers buy disks with less cache.

Suggesting that adding memory and/or flash to an enterprise array with the same number of spindles as a JBOD will not make it nearly half as slow is nonsensical.
Anonymous on Nov 17 2009, 22:11