Email thread on units used for storage

on July 13-14 2006

(compiled by Kors Bos for the Sept. GDB meeting at BNL)

Jeff Templon (NIKHEF)

Yo,

David Groep is in the process of designing our storage accounting system. The question of units came up. Hard disk manufacturers typically work in SI units, ie a terabyte is 10^{12} bytes or a million million bytes. I don't know how other Tiers did their MoU numbers, but we specified them also in SI units. I just checked the MoU and it makes no reference to the choice of units.

However, the unix system typically uses the "1024" scale. Apparently (no, I am not making this up) there is a new IEC standard for this, so your system should be reporting how many "kibibytes" your file is and not kilobytes, since kilobytes refer to the SI units and hence 1000 bytes.

We should clear this up now, because at the petabyte scale, the difference is 12% which might start to become significant.

I would personally prefer to stick with the standard SI units (based on powers of ten).

J "pebibyte sounds like what Pebbles might do to Bam-Bam" T

Andrew Samsun (RAL)

RAL has always reported in units of 1024 - both what we have pledged, made available and what the experiments have used. Reason being that the experiments told us that was what their requirement was measured in.

For available disk capacity we have quoted the storage capacity available to the experiments after all system overheads have been subtracted.

Does this mean I have just gained 12% extra capacity?

For tape capacity (to widen the discussion) we have quoted in units of 1024 and have used the average amouunt of data that Particle Physics stores on the tapes. This is a product of the physical capacity, the average occupancy and the average compression factor.

Regards
Andrew

Manuel Delfino (PIC)

Dear Jeff,

Indeed, 12% becomes significant for those of us who pay the bills.

In the case of planning for PIC, we used two things: Informal quotes from manufacturers (which for sure where in "SI" units) and the PASTA numbers as quoted in the CMS document CERN-CMS-NOTE-2004-031 or CERN-LHCC-2004-035 or LHCC-G-083 (http://cdsweb.cern.ch/search.py?recid=814248&ln=en) which was the only "official reference" we could find to the PASTA numbers at that time.

The question is, are the PASTA numbers in "SI" units? Bernd, are you on this list? I copy you in case you are not.
My preference would be to use "SI" units, since that is what the hardware orders and bills are based on.
Cheers, Manuel.

Andrew Samsun (RAL)

I would be glad to see a very clear and agreed process for providing information to LCG about both made available and used disk and tape capacity.

The amount of disk we purchase from suppliers has overheads in it that we don't include in our figures. For example not all our servers always contribute to experiment resource and are not included (we use a figure of 10% for our spinning hot spare reserve). Full RAID and filesystem overheads are included in our figures.

Regards
Andrew

Federico Carminati (ALICE)

Hello,
in our computing model we have used 1024 to move from one unit of xBytes to the next one. So if the decision is to use 1000, we have to revise our numbers. As Andres says, whatever decision is fine, provided it is well documented and generally agreed. A todo for the next GDB / MB? Ciao,

Federico Carminati

Bernd Pfanzer (CERN)

Hi Manuel,

yes, I am member of the GDB list.

Now I try to give a sort of answer to your question. Unfortunately it is not simple :-))

what defines the 'available' disk space ? :

1. the space on the disk itself is reported as billion bytes by the vendors
   500 Gbytes = 500 000 000 000 bytes

2. these disks will be 'aggregated' into a hardware RAID system.
   Now it starts to depend on the details of the RAID set-up and the details of the general node hardware
   e.g. disk server with 16 500 GB disks
   -- maximum space with one 16 port controller and 15 data disks and one 'parity' disk
      space = 15 * 500 GB
   -- a 16 port controller is more expensive than 2 8 port controller, but now one has to have 2 RAID systems as one can't span over two controller. Plus we want a bit more safety, thus we go for 6 data disks + one parity + one spare
      space = 12 * 500 GB
   -- all kinds of other variations (e.g. RAID1 for analysis performance...)
   -- during time and dependent on usage characteristics one might need to change the internal configuration of the 'same' disk server --> performance versus space versus 'reliability'

3. now we add a file system (ext3 or xfs) and we loose ~ 2 % of the total space per file system. The reported file system usage depends on the parameters of the df command, usually it is reported in 1k-blocks (1024 bytes).

4. how much of the space in a file system can one actually use :
   -- garbage collection setting of the governing mass storage system defines upper limits
   -- files get their space in an allocation of 4096 Byte blocks, thus a large
      number of small files gives a worse efficiency than fewer large files
   -- as soon as one has multiple writers and the file system gets close to 90-100%
      (depends on the # and the file sizes) fragmentation starts and one gets bad
      reading performance afterwards.
      The situation is dependent on the application :
      high frequency staging analysis pool versus high performance mass storage to tape pool
      versus 'organized' analysis write-once read-many pool
      -->HSM policies and load-balancing + experiment specific storage management system
      Thus in some cases one might want to stick to <90 % occupation

All together we can easily get >>10% 'available' space variations and it depends on the usage.

Anyway, this possible 10 % (1000 versus 1024, all our internal numbers are in TB) is within the error bar of any cost predictions for the next few years and the 'systematic' error (which I just described) is of the same magnitude and higher (not to mention any budget 'short-comings').

So, a proposal for space reporting could be :
-- we use 1 KB = 1000 Byte reference
-- reference is the amount of 2 Gbyte ($2 * 10^9$ bytes) files which could fit into the available
   disk servers (after file system creation and assuming 100 % fill rate)

The experiments gave their disk space requirements itself with an included efficiency of 70 %, but I am sure that this does not include the mentioned points.

Just a few thoughts about this interesting topic.....

cheers, bernd

Tony Cass (CERN)

Jeff,

CERN buys disks in units of 10**3N but reports in units of 2**10N. Given filesystem are built (and report) in units of 2**10N this seems the logical choice.

Cheers,
Tony

Federico Carminati (ALICE)

Hello,
RAL practice is consistent with the ALICE computing model. Any change to this would imply a change in our numbers (i.e. you did not gain 12% ;-) Best,

Federico Carminati

Andrew Samsun (RAL)

Its always depressing how the multiple factors of 1024, system overheads, spare systems etc etc hit the bottom line.

Manuel Delfino (PIC)

Hi Bernd, thanks for the detailed answer. Given this and the other messages, it seems my initial reaction was wrong.

I guess what we are talking about here is the "external" or "user perceived" storage space.

For actually used space, this is simply sum_over_all_files_on_given_medium(size_of_file)*normalization. By "given_medium" I mean "disk" or "tape" or "durable" or whatever. Given this, I could care less about what "normalization" is, as long as we all use the same.

For "allocated" or "installed" storage space, each center will have to derive a conversion factor from installed hardware to "user perceived" space, using their experience and the details of their setup, along the lines described by Andrew and Bernd.

If this is right, perhaps not much more discussion is needed, but it would be good to "officialize" it at the MB and agree on "normalization".

If this is wrong, please correct me!

Cheers,
Manuel.

Tony Cass (CERN)

Manuel,

Your "sum_over_all_files_on_given_medium(size_of_file)*normalization" strongly suggests to me that normalization should be 1. Users think of TB as TibiBytes, not TeraBytes and I don't think this will change soon; this is Federico's point. To be meaningful to the users, the allocated and installed space have to be in the same units.

The normalization is all in the price. It is not possible to translate the "MediaMarkt" price for a 500GigaByte disk directly into a cost for a PibiByte-scale storage service. The cost of such a service has to take into account all the factors discussed by Bernd. And sites will wish to trade off purchase cost against operational cost as well...

Tony

David Foster (CERN)

Of course if we are going to introduce new terminology we have to agree what that is first. Although I have not heard of TiBiByte, I had heard of TeBiByte (Tera Binary Byte) as meaning explicitly 2^{40} bytes as opposed to the decimal interpretation of TeraByte as 10^{12} Bytes although we all understood TeraByte as 2^{40} anyway ...

Similarly for GiBiByte (Giga Binary Byte) and PeBiByte (Peta Binary Byte) .....

Still, whats is a name :-)

cheers David.

Manuel Delfino (PIC)

Hi Tony,

I also think 1024 is the right thing for physicists. After all, this whole thing comes from the fact that BillG did not want to divide by 1024 in DOS.

I dont understand your second statement, though. By

sum_over_all_files_on_given_medium(size_of_file)*normalization

I simply mean that if size_of_file is in bytes, and we have agreed to work in 1024s, then we should divide by 1024**n to report the number. This is simply due to the fact that it is impractical to report numbers in bytes.

What you call "normalization" in "The normalization is all in the price." seems to refer to what I called the "conversion factor" (from installed hardware to "user perceived" storage) combined with earlier statements about how to calculate costs/prices.

Are we on the same wavelength? Keeping consistent terminology is important, and unfortunately "normalization" is in the same class of words as "control" -- they mean different things to different people.

Cheers,
Manuel.

Tony Cass (CERN)

Manuel,

OK, sorry; I hadn't understood your normalisation as dividing by 1024**n, but as multiplying by (1024/1000)**n or something, which I think Jeff was advocating in his original mail.

I'm now on the same wavelength.

Cheers,
Tony

Jeff Templon (NIKHEF)

Yo,

I am not sure I was advocating much of anything, I just wanted to know what we should use. I expressed a preference for SI (powers of ten) units, but that's only because I didn't want to have to start saying tebibyte without breaking into a laugh.

J "i can always try" T