LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)

Lossy Compression WG Report

A. Author, B. Author, and C. Author

DMTN-068

Latest Revision: 2018-04-17

DRAFT

Abstract

We report on the investigation into the use of lossy compression algorithms on LSST images that otherwise could not be stored for general retrieval and use by scientists.

Change Record

Version	Date	Description	Owner name
1	2017-04-17	Initial release. Based on LDM example	Tim Jenness
2	yyyy-mm-dd	Future changes	Future person

Contents

Lossy Compression WG Report

1 Introduction

The Lossy Compression WG was formed in response to RFC-325 with its charter being [LDM-582]. In RFC-325 it was recognized that user experience would likely be unacceptably impacted by the long latency required to access some LSST image data. Central to this concern is that the current data model does not support storage and serving of processed visit images (PVIs), i.e. the detrended, calibrated individual exposures from the survey. Instead, users needing such images would either have to rely on retrieval from tape media or regeneration of the PVIs on-the-fly.

Previous analysis has indicated that retaining all processed images on disk would be too costly and therefore not feasible, unless lossy compression is applied. The same analysis did indicated that storing all raw data on disk (with a loss-less compression) is feasible. The Lossy Compression WG was asked to investigate whether some pipeline products might be saved after applying a lossy compression algorithm without significantly degrading their suitability for a wide range of scientific investigations. Central to this is the need that the compressed products be small enough that the cost to store and serve these images could be met within a reasonable budget. The benefit from storing compressed products would only be realized if those products were indeed useful for many users as it would free resources that otherwise would be engaged in regenerating or serving a tape archive.

The LSST has traditionally avoided lossy compression for any of its image data products (including the large co-added images as well as templates retained for each data release). Anecdotal experience from other recent surveys indicate that science ready images stored with a lossy compression satisfy the scientific needs of their user communities. For example, the Dark Energy Survey (DES), uses FPACK (Pence et al., 2009) with a quantization of 16, and Pan-STARRS reportedly uses 4-bits per standard deviation (also equivalent to a quantization factor of 16) but with an inverse hyperbolic sine transformation to concentrate the sampling near the background level (Waters et al., 2016). Indeed, Price-Whelan & Hogg (2010) have argued that none of the scientific information is lost in an astronomical image even with fairly drastic quantization (at levels as high as 0.5σ). The tests used by Price-Whelan & Hogg (2010) were relatively idealized, in this note we describe the results from a small test using precursor data from HSC to provide a sense of how lossy compression might be applied for LSST.

2 Methodology

This investigation is not meant to address the specific file format(s) that might be used to store LSST data (e.g.; FITS vs. HDF5). The tests that have been made were performed using images stored using FITS, mainly because the changes necessary could be used within the current LSST pipeline testing infrastructure. The specific images used were a set of HSC data that formed a modest depth patch on which pipeline regression testing was already being routinely performed in the development of the LSST pipelines (the *ci_hsc* test set). For this test set there were 33 images/CCDs, from 11 visit/exposures, at two bands (HSC-R, HSC-I). Included among these images are a 4 images near the edge of the HSC focal-plane, where vignetting causes a portion of the detector unusable for science. These regions are masked and present very different noise characteristics but are useful because they show some caveats that must be considered when applying compression.

As part of this investigation we have separated the loss from the actual compression algorithm. A change has been injected into the pipeline that allows for a quantization to be applied to the science (and weight) images that are traditionally stored as floats. Formally, the quantization factor, q, determines the number of samples/subdivisions of some set number, in this case the standard-deviation of the image pixel values that do not contain a detected source. For a FITS image this is expressed as a scale factor (BSCALE) and the image pixel values are converted to the nearest integer multiple of this factor. We then use existing loss-less compression algorithms to compress the integer representation of the image to achieve a compressed image. Our tests varied the factor q from 4 to 128 (stepping by factors of 2).

Metrics are then obtained to understand the impact and efficacy of compression. Broadly, these fall into three categories:

1. **Image Compression benchmarks:** to measure the changes at the pixel level. These include: percent increase in noise/RMS, median difference, and number of pixels that change by more than the quantization level (to catch cases where the integer representation is not able to capture the full dynamic range of the original images).
2. **Catalog/Measurement benchmarks:** to measure the change of aggregate quantities of interest for scientists using the images for scientific measurements. The current benchmarks being measured are source position, flux, and shape along with their associated uncertainties.
3. **Compression algorithm benchmarks:** to measure the compression factor achieved,

along with algorithm execution times for compression and decompression.

In addition a second round of image and catalog benchmarks can also be obtained to assess the changes that might be expected when compressed products are combined to form stacked images from which astronomical source measurements are also obtained.

3 Results

3.1 Single Image Compression Benchmarks

At the image level, independent measurements of the noise in the original science and weight images (I_0, W_0) and the quantized versions (I_q, W_q) are made. The algorithms used are independent of those that performed the estimates used to set the quantization. In most cases we consider only pixels with FLAG=0 or FLAG=32 (which indicates the presence of a source) as heavily masked regions often have values (particularly in the weight image) that can exceed the range accessible in the quantized images. Figure 1 shows the distributions of pixels values for the science and weight planes from two images typical of those in the test set.

A base level check is made that examines the difference between the quantized and unquantized version of an image ($I_{diff} = I_q - I_0$). First the mean, $\bar{I}_{diff}$, and RMS, $\sigma_{I_{diff}}$, are computed to show that no systematic offset occurs and that the noise in the difference is indeed less than the scale factor. We then also search for pixels where the difference exceeds the quantization level. For most images this latter value is identically zero but in a small number of cases the pixels in a bright object will exceed the range available in the quantized image (i.e. the integer representation has insufficient cardinality to track the dynamic range in the image). If flagged pixels are included, then are typically more pixels that exceed this range and in the worst cases (e.g. images from CCDs that are vignetted) a large fraction of the weight pixels cannot be tracked. **More needed to quantitatively describe these?**

We then measure the standard deviation (RMS) in each science and weight image (σ_{I_q} and σ_{W_q} respectively) to understand the fractional increase in the image noise from the quantization ($\sigma_{grow} = \sqrt{\sigma^2_{I_q} - \sigma^2_{I_0}}$. Figures 2 and 3 show histograms of these metrics based on the images in this test set. The left panels show the residual noise as measured from the difference between the unquantized and quantized images. The right panels show the fractional additive noise resulting from the quantization. Note that the number of samples in the histograms for q=64 and 128 are smaller than the total because the measurement of the standard deviation is

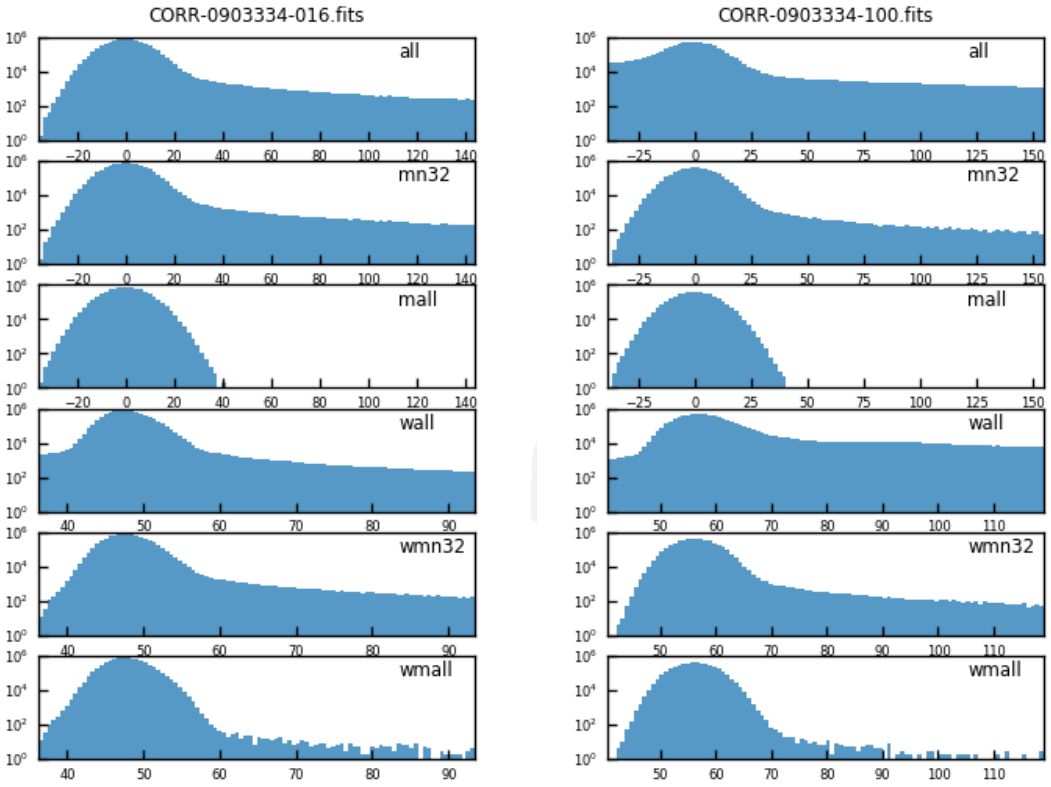

FIGURE 1: Distributions of pixel values from two images in the test set. The left set of panels show distributions for a "normal image", while the right panels show the distributions for an image near the edge of the focal plane with heavy masking. (top) to (bottom) the panels show: all science plane pixels (all), all science pixels with MASK=0 or 32 (mn23), and all unmasked pixels (mall), followed by similar distributions for the weight plane (wall, wmn32, and wmall). Note the "mall" and "wmall" are roughly the distribution that was used to estimate the quantization level.

TABLE 1: Differences in Coadd Images Constructed from Quantized Images

q	med(δ_I)	std(δ_I)	min(δ_I)	max(δ_I)	med(δ_W)	std(δ_W)	min(δ_W)	max(δ_W)
Sample 1 (HSC-R coadd)								
q4	3.6×10^{-5}	1.3×10^{-2}	-1.1×10^{-1}	9.0×10^{-2}	$< 10^{-6}$	3.8×10^{-5}	-5.7×10^{-4}	5.7×10^{-4}
q8	-4.0×10^{-6}	6.6×10^{-3}	-5.3×10^{-2}	4.2×10^{-2}	$< 10^{-6}$	1.9×10^{-5}	-2.8×10^{-4}	2.8×10^{-4}
q16	-1.1×10^{-5}	3.3×10^{-3}	-2.4×10^{-2}	2.1×10^{-2}	$< 10^{-6}$	1.0×10^{-5}	-1.4×10^{-4}	1.4×10^{-4}
q32	3.0×10^{-6}	1.7×10^{-3}	-1.4×10^{-2}	1.4×10^{-2}	$< 10^{-6}$	5.0×10^{-6}	-7.1×10^{-5}	7.1×10^{-5}
q64	6.0×10^{-6}	8.3×10^{-4}	-6.5×10^{-3}	5.6×10^{-3}	$< 10^{-6}$	2.0×10^{-6}	-3.6×10^{-5}	3.6×10^{-5}
q128	2.0×10^{-6}	4.1×10^{-4}	-3.4×10^{-3}	3.4×10^{-3}	$< 10^{-6}$	1.0×10^{-6}	-1.8×10^{-5}	1.8×10^{-5}
Sample 2 (HSC-I coadd)								
q4	4.4×10^{-5}	2.3×10^{-2}	-1.7×10^{-1}	1.7×10^{-1}	$< 10^{-6}$	8.0×10^{-5}	-1.1×10^{-3}	1.1×10^{-3}
q8	-3.8×10^{-4}	1.2×10^{-2}	-8.9×10^{-2}	8.4×10^{-2}	$< 10^{-6}$	4.0×10^{-5}	-5.6×10^{-4}	5.6×10^{-4}
q16	-3.4×10^{-4}	5.9×10^{-3}	-4.3×10^{-2}	4.6×10^{-2}	$< 10^{-6}$	2.0×10^{-5}	-2.8×10^{-4}	2.8×10^{-4}
q32	-3.3×10^{-4}	3.1×10^{-3}	-2.5×10^{-2}	2.5×10^{-2}	$< 10^{-6}$	1.0×10^{-5}	-1.4×10^{-4}	1.4×10^{-4}
q64	-3.3×10^{-4}	1.8×10^{-3}	-1.8×10^{-2}	1.7×10^{-2}	$< 10^{-6}$	5.0×10^{-6}	-7.0×10^{-5}	7.0×10^{-5}
q128	-4.0×10^{-6}	7.3×10^{-4}	-5.9×10^{-3}	5.2×10^{-3}	$< 10^{-6}$	2.0×10^{-6}	-3.5×10^{-5}	3.5×10^{-5}

approaching the machine accuracy (i.e. σ_{I_q} differs from σ_{I_0} by less than a part in 10^6).

3.2 Composite Image Benchmarks

Beside the individual images, the current tests construct a coadded patch. Here, we compare the resulting coadd images constructed from the original, never-compressed images with coadd images constructed from the quantized images. In the current limited test, only two coadded images (patches) were produced. The comparison is further hampered both because the depth of these coadd images is shallow (there are only 5 or 6 visits being combined per coadd) and an outlier rejection algorithm is active within the pipeline. This comparison is similar to that made for the individual images except that a a constraint has been added to remove locations where the clipping algorithm has systematically rejected a region of an image (in the test set there were of order a few such regions per coadd image totaling comprised of a few times 10,000 pixels at q=4 dropping to 1,000 pixels at q=64).

The results are summarized in Table 1 which shows.... the

3.3 Catalog/Measurement benchmarks

Here we outline the comparison of measurements made on individual ccd-visit images with and without quantization applied. Currently four types of measurements are considered: aperture photometry, PSF photometry, centroids, and shapes. In each case the comparison

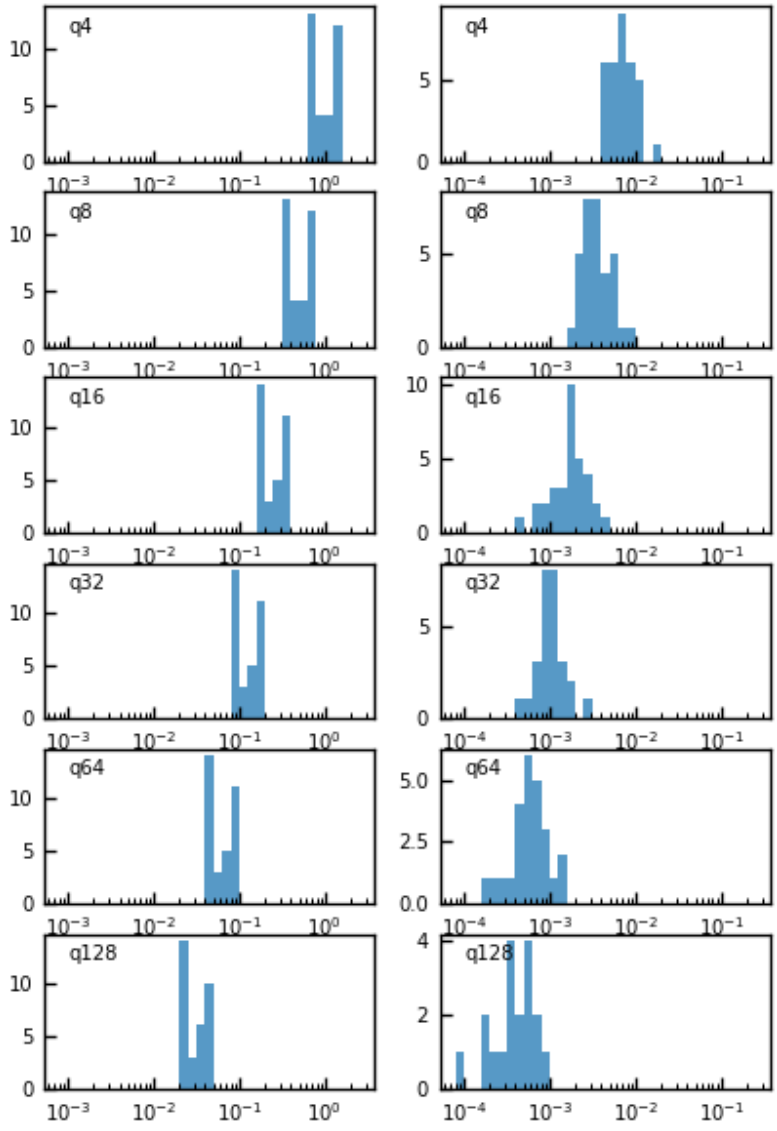

FIGURE 2: Histograms showing image level statistics with respect to the original compressed image. (left panels) are histograms showing the RMS of the difference between the compressed and original image. (right panels) are histograms of the fractional increase in the noise with respect to the original image.

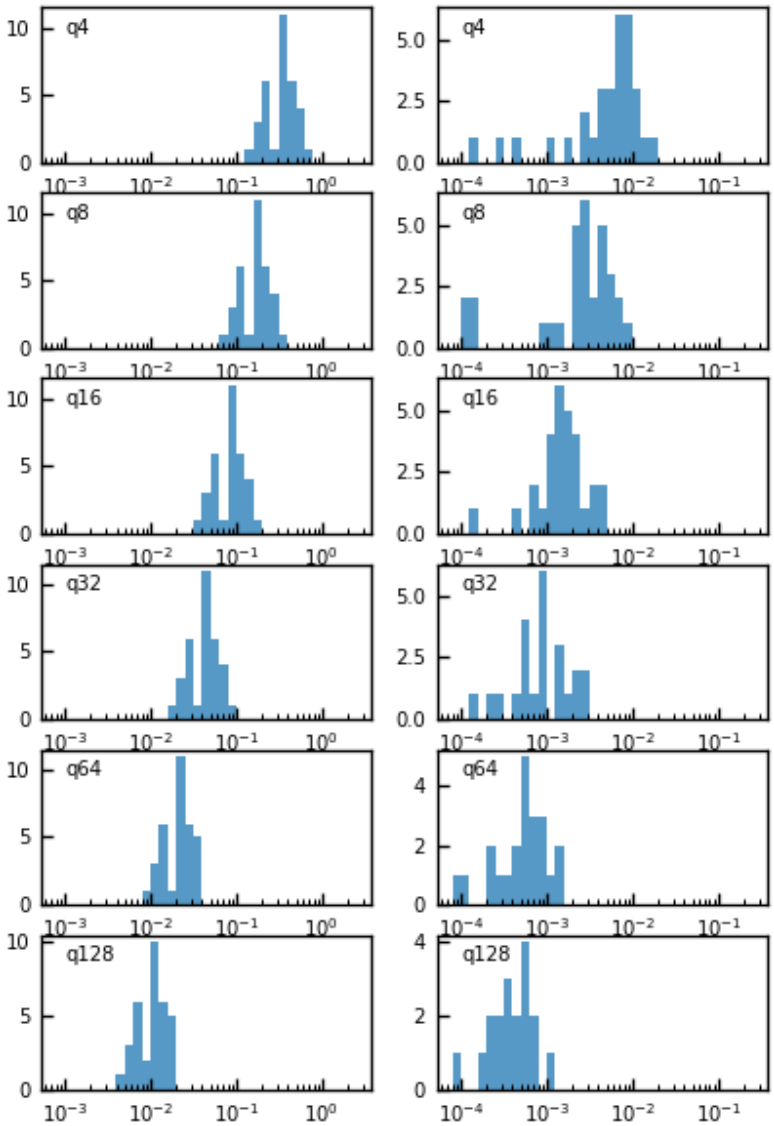

FIGURE 3: Similar to Figure 2 but for the weight image.

is made by using forced photometry based on the COADD catalogs from the *ci_hsc* run without quantization. An astrometric match is made between the catalog from the never-quantized images to each of the catalogs from the quantized images with a 1'' match radius (with the nearest source being considered the match). The results from multiple CCDs are accumulated into a single plot in order to obtain statistics at the bright end.

Figures 4 show comparisons for flux measurements for aperture photometry and PSF fitting. The aperture photometry measurements *base_CircularApertureFlux_6_0* use a 6 pixel radius circular aperture while the PSF fitting measurements are the *base_PsfFlux_flux* measurements. Note that in these plots no star-galaxy classifier was used to subselect stellar/point-source measurements.

The top two panels in each set show the total number of objects per flux bin, followed by a plot showing the flux uncertainty as a function of flux from the never compressed image. Beneath these are plotted the difference between the measurements from the quantized images and the never-quantized images with subsequent plots using an increasing level of quantization. These difference plots are shown in units of σ_{F_0} (i.e. each difference measurement is scaled by the uncertainty in the flux measured in the unquantized image). Overplotted are histograms showing the difference level that encompasses 50, 75, 90, and 99% of the measurements as a function of flux bin. **Remove 99% histogram?** Careful examination of the histograms show that for quantization, q=16 or greater 50% of all flux measurements differ by less than 0.01σ and 90compared to the measurements on the unquantized images.

Similar to the flux measurements, Figure 5 shows comparisons of centroid and shape measurements (left and right panels, respectively) as a function of signal-to-noise (S/N) in the unquantized images. For the centroids we use the *base_SdssCentroid_x* (x), and *base_SdssCentroid_y* (y), to compute the linear offset $X_q - X_0 = \sqrt{(x_q - x_0)^2 + (y_q - y_0)^2}$ between the measurements made in the quantized and unquantized images. Note that the current version of forced photometry does not flag poor and low signal-to-noise measurements so those measurements pollute/inflate the distributions show in the low signal-to-noise portion of the centroid plots in Figure 5.

In order to investigate the impact of quantization on shapes, we use the *base_SdssShape_xx*, *base_SdssShape_yy*, and *base_SdssShape_xy* measurements to form a shape measurement, S, where $S = (I_{xx}I_{yy} - I_{xy}^2)^{1/4}$. Assuming that those 2nd moment measurements are not strongly correlated, we also define the uncertainty in S as $\sigma_S^2 = (\frac{\partial S}{\partial I_{xx}})^2\sigma_{I_{xx}}^2 + (\frac{\partial S}{\partial I_{yy}})^2\sigma_{I_{yy}}^2 + (\frac{\partial S}{\partial I_{xy}})^2\sigma_{I_{xy}}^2$, and use the associated uncertainties to estimate σ_S. Figure 5 shows these measurements and

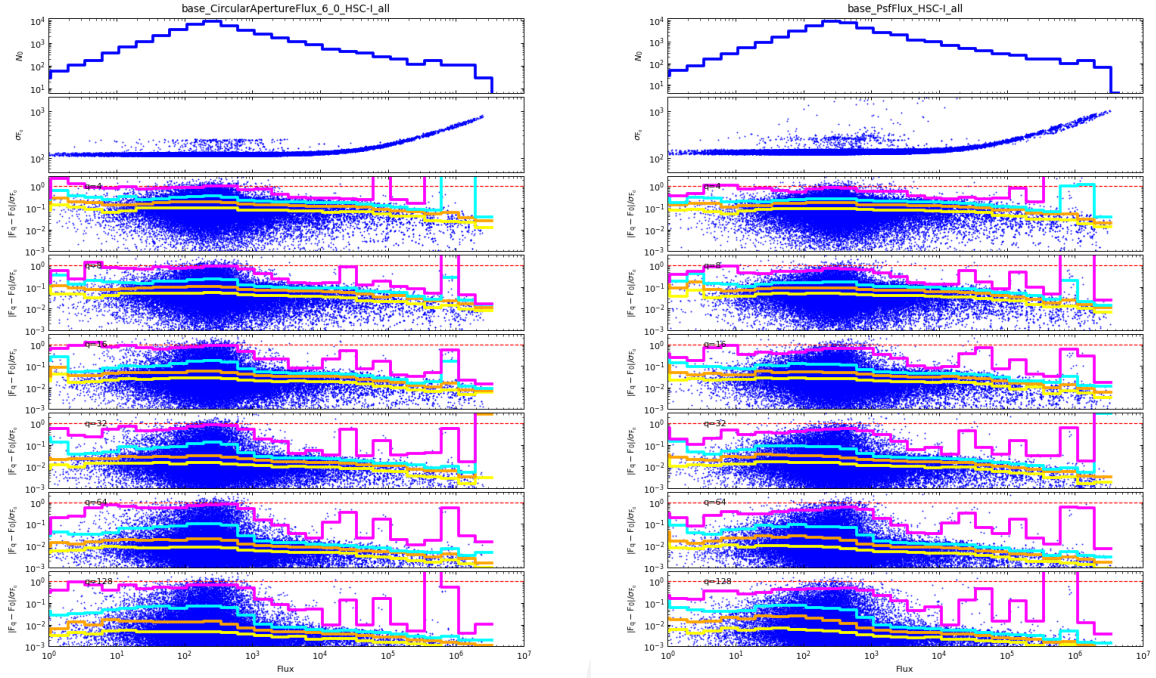

FIGURE 4: Comparison of aperture (left) and PSF photometry (right) measurements resulting from forced photometry on individual images with and without quantization/compression. The top panel in each shows the distribution of objects as a function of their flux measured in the original image(s). The second plots show measure uncertainties for those flux measurements. The panels below show the difference between the flux measurements made on the unquantized and quantized images divided by the uncertainty in the quantized images (in units of σ_{F_0}. A dashed horizontal red line shows the 1σ difference level for reference. The lower panels are for measurements from the images with progressively higher quantization factors (less loss). The histograms in each panel show the difference level at which 50, 75, 90 and 90% of the objects are found.

makes a comparison of the quantized measurements to those found with the unquantized images. Measurements with *base_SdssShape_flag* have been excluded.

Examination of the centroid plots in Figure 5 show that for quantization of q=16 or better the difference in centroid measurements is less than 0.1 pixels for 90% of measurements for sources with S/N>10. Similarly, for shape measurements the differences in $\mathcal{S}$ are already less than 0.01 pixels for 90% of objects at a S/N>10.

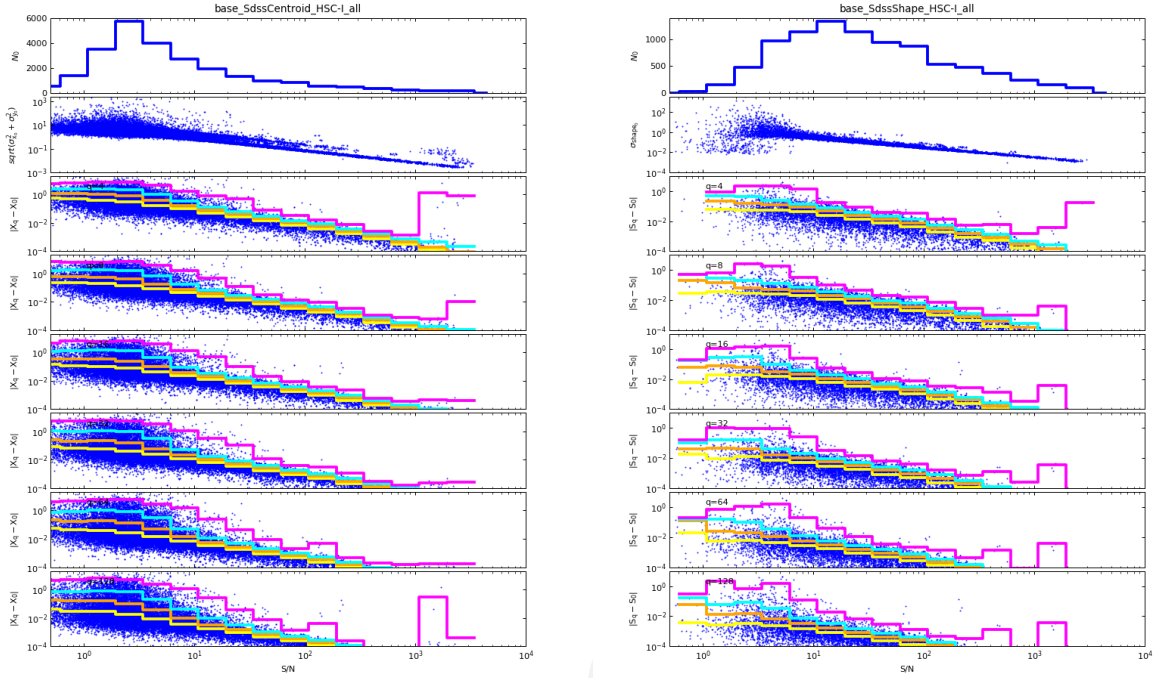

FIGURE 5: Similar to Figures 4 but for centroid (left) and shape (right) measurements as function of the signal-to-noise ratio of the unquantized measurements. The differences in the lower panels are in units of pixel offset and pixel radius for the centroid and shape measurements, respectively.

3.4 Catalog/Measurement from COADD images constructed from quantized PVI images

Similar to the comparisons made for the individual images, we compare the catalog measurements from the coadded patch that was constructed from the never-quantized and the quantized PVI images. Note, no further quantization/loss was applied to the COADD images. The same four quantities were examined (aperture flux, PSF flux, centroid, and shape). Figures 6 through 7 show the results of that comparison.

3.5 Compression Algorithm benchmarks

We have applied a variety of existing compression algorithms to the quantized images from this study to obtain benchmarks of their efficacy. The values reported reflect those algorithms' performance when running under OS X 10.13.2 (macOS High Sierra) on a MacBook Pro with

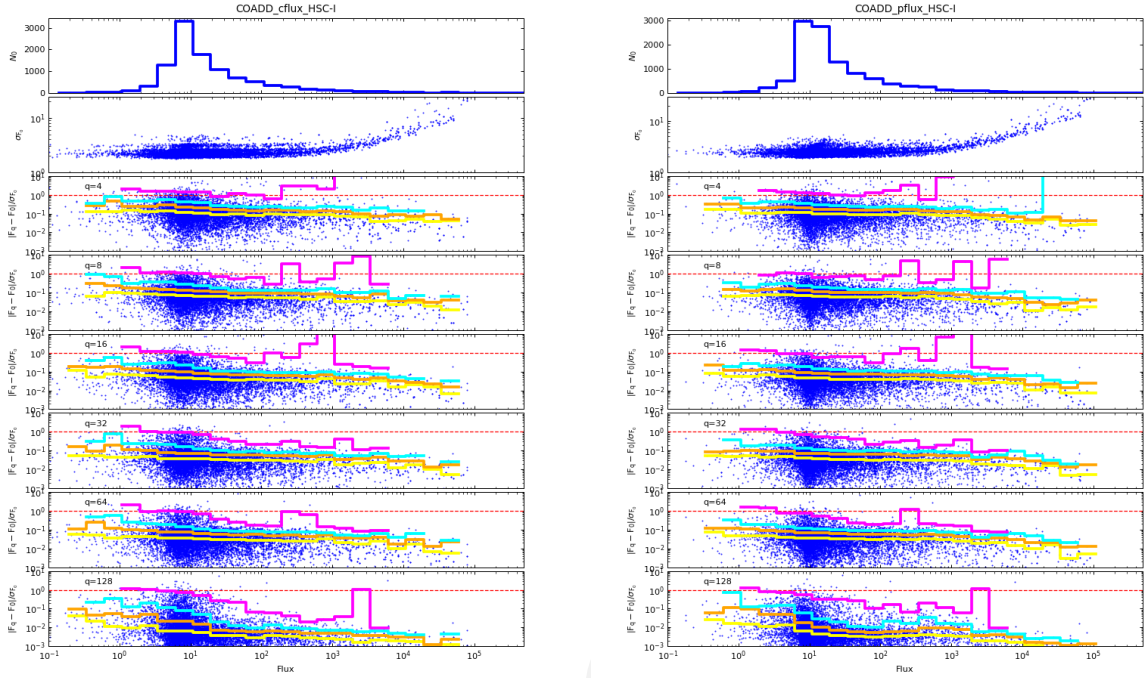

FIGURE 6: Similar to Figure 4 but for measurements made on the coadd image.

quad 2.9 GHz processors. A ramdisk was used for storage to minimize the impact of I/O operations within the test.

A range of existing algorithms have been benchmarked, including a number which use threading to achieve greater speed. The algorithms considered were:

1. **gzip:** the standard GNU implementation of Lempel-Ziv (LZ77).
2. **pigz:** a threaded version of gzip.
3. **bzip2:** an implementation of Burrows-Wheeler block sorting (offers the possibility of recovery of undamage block).
4. **pbzip2:** a threaded/parallel implementation of bzip2.
5. **lbzip2:** another threaded/parallel implementation of bzip2.
6. **lz4:** a ”typically faster” implementation of LZ77 (favoring speed over compression ratio). Pushing to higher compression ratios significantly degrades performance.

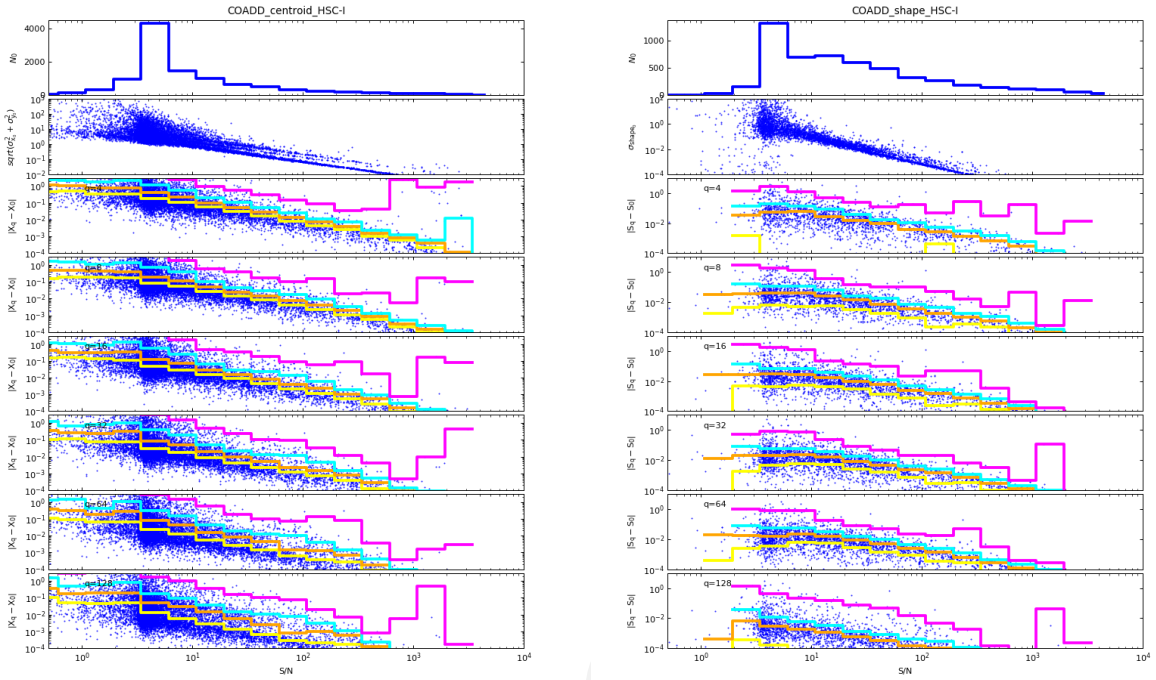

FIGURE 7: Similar to Figure 5 but for measurements made on the coadd image.

7. **lzop:** a separate implementation that trades a small hit in compression time for an improvement in decompression. The performance trade is not apparent at the file size being used in these tests.

8. **zstd:** Also based on the LZ77 family and includes a parallel implementation. In addition this algorithm has implementations/bindings over a wide variety of languages (including Python). The usage of a pre-computed dictionary may offer improvement in speed/compression factor but rigorous testing was not possible for this small set (performance was identical to the untrained algorithm if the full set was used both to train and then obtain benchmarks).

9. **xz:**

The results from benchmark tests are summarized in Tables 2-4, showing compression factor, time to compress per file, and time to decompress per file, respectively. These times do not include the time to necessary to obtain and apply scale factor used in the quantization. Furthermore, the set of files being compressed are nearly identical (98 Mb) and therefore do

Table 2: Compression Factor Achieved

q	gzip	pigz	bzip2	pbzip2	lbzip2	lz4	lzop	zstd	zstdb
q4	6.73	6.73	9.96	9.95	9.96	3.69	3.11	6.29	6.29
q8	5.54	5.53	8.20	8.20	8.21	3.34	2.96	5.42	5.42
q16	4.69	4.69	6.81	7.01	7.03	3.11	2.82	4.82	4.82
q32	4.04	4.03	6.14	6.14	6.14	2.93	2.66	4.35	4.35
q64	3.62	3.62	5.47	5.47	5.48	2.82	2.47	3.94	3.94
q128	3.38	3.37	4.88	4.88	4.88	2.66	2.32	3.56	3.57
vanilla	1.71	1.71	1.80	1.80	1.80	1.50	1.49	1.72	1.72

Table 3: Time to Compress per File

q	gzip	pigz	bzip2	pbzip2	lbzip2	lz4	lzop	zstd	zstdb
q4	4.45	1.18	5.00	1.42	0.85	0.21	0.24	0.36	0.12
q8	6.06	1.64	4.91	1.39	0.82	0.21	0.24	0.42	0.15
q16	8.27	2.24	4.33	1.39	0.82	0.27	0.27	0.55	0.18
q32	10.30	2.76	5.27	1.42	0.79	0.24	0.27	0.58	0.21
q64	11.79	3.00	5.39	1.52	0.88	0.24	0.30	0.61	0.24
q128	12.76	3.21	5.91	1.61	0.94	0.27	0.30	0.67	0.21
vanilla	3.36	0.97	8.94	2.79	1.58	0.15	0.12	0.30	0.15

Table 4: Time so Decompress per File

q	gzip	pigz	bzip2	pbzip2	lbzip2	lz4	lzop	zstd	zstdb
q4	0.21	0.24	2.30	1.21	1.27	0.15	0.18	0.27	0.24
q8	0.24	0.27	2.33	1.12	1.24	0.18	0.18	0.27	0.27
q16	0.27	0.27	2.02	1.12	1.21	0.18	0.18	0.27	0.24
q32	0.30	0.30	2.42	1.24	1.24	0.15	0.18	0.27	0.24
q64	0.30	0.30	2.42	1.27	1.09	0.18	0.21	0.27	0.27
q128	0.30	0.33	2.82	1.30	1.24	0.24	0.21	0.30	0.27
vanilla	0.39	0.36	4.36	1.48	1.27	0.15	0.12	0.24	0.24

not provide any information about algorithmic performance with respect to file size. When a parallel implementation was available the threading was set to use 4 cores.

4 Recommendations

Below, our recommendations assume:

- The capability to recompute a reduced-calibrated image product on-the-fly will be possible for users that need such.

- Astronomers have the scientific acumen to understand that measurements and products made using lossy-compressed images will not exactly match those made during release production.
- The tests in this note are inadequate in a couple of respects. The measurement algorithms are not those that will be deployed in the LSST Alert and Data Release Processing. The COADD images/catalogs in the current tests are comprised of a small amount of data and therefore cover a small area with shallow depth. In order to have a detailed understanding of the impact of compression a much larger dataset than a *ci_hsc* would be needed.

With these in mind we recommend the following:

1. **Quantization Factor:** Most scientific use cases should be satisfied by a quantization factor of q=16 or q=32. This would provide XX, YY, ZZ as summarized in Table **Work for Tuesday 17th**. **(Note this is being addressed now see #2 in Section 5 along with other comments being considered.)**
2. **Algorithm:** The best performing, off-the-shelf candidate for compression is BZIP2 which achieves a compression factor of 5-7. Its main drawback is speed but in trading speed for compression factor the use of LZ or ZSTD would roughly double the storage costs.
3. **When and where to use lossy compression:** Compression should occur after production but before archiving. This ameliorates any risk that compression adversely impacts the ability of LSST to meet science requirements. The availability of compressed products for users is meant to allow investigations to proceed without an explicit need for reprocessing.
4. **PVI Products:** PVI images are the clearest case where lossy compression should be considered as these products otherwise would not be stored, requiring re-computation, or would require a large tape storage infrastructure.
5. **Other Products:** Within the data storage model there are a few other products that might be considered as candidates for lossy compression. These are: the DRP COADD images, the AP templates, and the 60-day store of PVI images from the AP pipeline for PreCovery of transients. The realized benefit of storing any of these with lossy compression is 100's of times smaller than the DRP PVI images. Moreover, if lossy compression were used for any of these data types, there would be a direct impact on the production

results. Therefore we do NOT recommend use of lossy compression without a demonstration that its use would not prevent reaching survey requirements. To do so requires a detailed test with working versions of the pipelines and real? LSST data.

6. **Verification:** The current tests are not realized within the LSST framework or QA effort and they are moderately costly to make (they require production to be repeated for each level of quantization). It is recommended that a means to implement tests similar to those detailed here be considered so that as the pipelines and LSST measurement algorithms mature...

5 TBD, Comments, etc...

The following are the results of comments, suggestions to improve complete this analysis (currently from the WG). Each of these are being considered (and if others have suggestions I am willing to listen) but this effort has now reached the point where it is poised to embrace the enemy of the good (i.e. the perfect) so higher powers may be encouraged/invoked to truncate this list.

1. ADD: number of sources lost as of function q
2. ADD: Table with benchmark vs q for each set of measurements
3. ADD: Contribution due to Lossy compression in an error budget (look at delta-uncertainty vs q?)
4. ADD: Code in repository
5. **Done** ADD: explain separability between compression and quantization
6. **Done** ADD: PanStarrs: used compression on-the-fly with no complications or complaints... (4-bits per std q=16)
7. FITS compatibility....

6 WG Membership

Membership of roughly four people is optimal and should include persons familiar with weak-lensing and difference imaging concerns. The proposed membership is:

- Robert Gruendl (NCSA; **Chair**),
- Paul Price (Princeton),
- Bob Armstrong (Princeton),
- Krzysztof Findeisen (UW; replacing John Parejko),
- Sophie Reed (Princeton),
- Eric Morganson (DES/NCSA; observer)
- Ben Emmons (EPO Tucson; observer)

References

[LDM-582], Juric, M., Gruendl, R., 2017, *Losst Compression Working Group Charge*, LDM-582, URL https://ls.st/LDM-582

Pence, W.D., Seaman, R., White, R.L., 2009, PASP, 121, 414 (arXiv:0903.2140), doi:10.1086/599023, ADS Link

Price-Whelan, A.M., Hogg, D.W., 2010, PASP, 122, 207 (arXiv:0910.2375), doi:10.1086/651009, ADS Link

Waters, C.Z., Magnier, E.A., Price, P.A., et al., 2016, ArXiv e-prints (arXiv:1612.05245), ADS Link