Computing Resources for DES Weak Lensing

DES Weak Lensing Working Group

1 Executive Summary

We propose to support the DES weak lensing science effort by building a computing base at Brookhaven National Laboratory (BNL).

Lensing measurements are particularly computationally intensive, and make use of essentially all of the basic data products, from pixels to catalogs. The lensing signal and systematic effects are both subtle enough that probing them will require processing a significant fraction of the full DES data set. However, effective development requires a reasonably tight feedback loop between development and data processing. The need to process large amounts of data quickly enough to provide meaningful feedback to developers can only be met if significant computing is available.

On top of this, the lensing working group is developing multiple pipelines in parallel, which requires correspondingly more computing power but notably not more infrastructure. It makes sense to share computer resources and local expertise to aid all pipeline development.

Computer hardware purchases of $30,000/year for five years will meet these needs. The computing will be hosted at the RHIC ATLAS Computing Facility at BNL (RACF). The RACF will provide power, cooling, installation, and maintenance at zero cost, and we will receive $\sim$40% bulk discounts by purchasing alongside larger experiments.

A number of DES weak lensing group participants are already using a small compute cluster at BNL for science code development. Erin Sheldon of BNL and Mike Jarvis of UPenn have been developing the primary DES WL pipeline at BNL and all major tests runs of the code have been performed there. Zhaoming Ma is a postdoc at BNL and has developed and tested a new, highly computationally intensive method for PSF interpolation using Principle Component Analysis (PCA). He will run this code as needed throughout the survey. Mandeep Gill of Ohio State has begun working on an alternative WL code at BNL. Tim Eifler has begun running is two-point lensing code on the cluster. Also, Joerg Dietrich of Michigan now has an account and will begin testing his WL code on the BNL cluster in winter 2010. Any other interested parties are encouraged to join this effort at no cost to them. Tom Throwe, Brookhaven physicist and computer systems expert, will assist DES users in solving computing issues that arise. Funding for computing resources at BNL will ensure this development can proceed efficiently and at minimal cost.

2 Outline of Goals

The goal of the DES weak lensing working group (WLWG) is to support DES weak lensing science. This science relies on many data products from the survey, including the images, calibrated fluxes, astrometry, as well as derived products such as PSF characterization, galaxy shear estimates, and galaxy photometric redshift distributions.

The WLWG will primarily support the science through development of pipelines to derive accurate shear estimates. These measurements require touching all the pixels, and so are computationally intensive. The group is also developing multiple pipelines in parallel in order to converge on an optimal method and for consistency checks.

3 Current Algorithms and the Development Cycle

The pipeline currently developed by Mike Jarvis and Erin Sheldon is incorporated into the DESDM. The pipeline is run on data produced by the DESDM at earlier stages in the processing. These are images that have instrumental signatures removed and basic catalogs generated from those images. The DESDM will run the weak lensing pipeline on a time scale corresponding to the yearly data releases.

For efficient development of the algorithms, many more processings will be required. Weak lensing methods are still evolving, and there is much work to be done in order to produce shear estimates sufficiently free of systematics to reach our science goals.

Supporting this research will require significant computing resources. For DES science we must measure one percent shear signals to an accuracy of ≈ 1 part in 100. But the noise per galaxy shear estimate due to the intrinsic shape of the galaxy is of order 30%, so one must reduce a large number of galaxies in order to even test the pipeline to desired accuracy. In order to characterize the signals and systematics to the required level under a wide variety of observational circumstances, a significant fraction of the total data set must be processed, which in turn requires computing.

In addition, a tight feedback loop is required between data processing and development. The only way to process such huge amounts of data quickly enough to provide sufficient feedback is to assemble a large amount of computing.

Once the reduction pipeline has run to create the galaxy shape catalog from the pixel data, we must also analyze the catalog data to extract parameters of the dark matter and dark energy. Some of the most powerful of these analyses are computationally intensive. In the next section we will give some examples of computation times in existing datasets.

4 Example Timings for Weak Lensing Codes

4.1 Basic Pipeline Timings

Table 1 gives some example timings on DC4 data using the pipeline developed by Jarvis and Sheldon. For these timings we used the small astro cluster at BNL. The astro cluster, at that time, had three compute nodes, each with 8 Intel Xeon 3GHz cores and 32GB RAM.

Table 1. DC4 Timing Numbers on the Bach Cluster at BNL

Data Set	# images/tiles	Memory Per Job	CPU hours
DC4 SE	46,500	1G	1728
DC4 ME	22,402/224	15-48G	576

Note. — Resources used for processing *i*-band DC4 images using the astro cluster at BNL in 2009. The Bach cluster at that time consisted of three compute nodes with 8 cores and 32G RAM each. DC4 SE is all the SE images (4k by 2k chips) available at the time DC4 was released. DC4 ME is the multi-epoch data: Catalogs derived from the coadd tiles and all the corresponding SE images that contributed to each tile. Only the unique images are reported in the count. Note the timings for coadd ME analysis would be significantly longer if the astro machines did not have high memory.

These are timings for the weak lensing pipeline alone: the images and catalogs used as input are generated beforehand by the DESDM.

In DES each bit of sky will be imaged during multiple epochs. Thus there are generally two types of algorithms for measuring shear: those processing a single epoch (SE) and those simultaneously processing multiple epochs (ME). In table 1, the SE data set is every image we had available at the time DC4 was released. This may include more images than the "official" DC4 release. When multiple processings of the same image were available, we used the newest. The ME analysis made use of a subset of these images, about half.

One "image" here means an individual 4k by 2k CCD exposure. "Tiles" are combined "coadd" images of all exposures covering a predetermined area of the sky. There are a number of steps in the SE processing. We find bright stars, characterize the PSF, interpolate the PSF to the location of all objects, and finally determine a best shear for each object based on associated pixels and PSF. For our current code, this takes about 160 CPU-seconds per image. The processing of each image is entirely independent and is currently using 8 cores in parallel for each image. Further parallelization is trivial.

For ME processing, we select objects from the coadd catalogs. We then transform the coordinates back into the individual SE images that contributed to the coadd. For each of these SE images we reconstruct the PSF as determined during the SE processing described above. We then perform a joint fit for the shear across all images. This takes about 2.5 CPU-hours per coadd tile. The code uses all 8 cores in parallel. This works out to be about 90 CPU-seconds per SE image contributing to the tile. Processing each tile is independent but depends upon the previous SE processing to get PSF information.

Note the memory usage for the ME processing is very high, ranging from 15G to 48G depending on the number of SE images that contribute to each coadd tile. The average

is about 20G. The machines in the Bach cluster have 32Gb memory, so all but a few tiles fit into memory. The processing is significantly slower when the machine has less than the required memory; e.g. 30-40% slower if the computers had 4G instead of 32G.

4.2 Analysis Timings

Many lensing analyses are relatively quick, but some require significant computing time. For example, the cluster mass and luminosity analysis presented in [1] took two weeks running on a 300 processor cluster. The computers used were about a factor of two slower than the CPUs in the Bach cluster. DES data set will be 50 times larger.

The three-point shear function and the PCA PSF decomposition are also computationally intensive, each taking of order a week to run on the fiducial compute system we propose below. But these analysis will need to run dozens of times: For the PCA we must explore the quality of the interpolation for different algorithmic models for the spatial variations and the number of important principle components. For the three-point function, we expect to recalculate this for different combinations of shears from various redshift bins.

5 Extrapolation and Requested Resources

5.1 Timing Extrapolation

The full DES survey is expected to generate 80,000 exposures over five years. Each exposure generates 62 CCD images, for a total of 5×10^6 images. The on disk data volume required for lensing analysis will be $\sim$100 TB of disk space when stored in compressed format, and including the data quality indicators for each pixel, plus the coadded images and catalogs that are produced by the DESDM pipeline and serve as inputs to the ME process.

We require a further factor of three in disk for redundancy in the distributed file system (Hadoop[1]). We use a distributed file system in order to get the required data throughput; we have found that a few dedicated file servers cannot handle the load. The redundancy guarantees data integrity and dramatically increases throughput. Distributed file systems have been in use in high energy physics and industry for many years, and have proven to be a simple, high-availability and high througput solution.

Scaling from the hardware and software speeds used for table 1, the SE processing is simply (1728 cph)/(46000 im)*(80000 exp*62 im/exp) gives about 21 cpu-years. The multi-epoch processing is more difficult to scale, but assuming the relative number of images per tile is similar in the future, we predict (576 cph)/(22402 im)*(80000 exp)*(60 im/exp) gives 14 cpu-years. On an 120 node cluster of equivalent 2010 machines this could be processed in 2 weeks. This is approximately the desired processing time given that we will want to reprocess the data with several iterations on two or more independent shape algorithms in order to obtain a robust result.

We can expect to speed up the current algorithms but also should be prepared for future algorithms that require more computation. We will also gain as processing power follows Moore's law. This gain has traditionally been in transistor density on a single device, but

[1] http://hadoop.apache.org/

Table 2. Projected Computing Purchases

Fiscal Year	Disk Storage [TB]	$ for Storage	Compute Servers 2010 Equivalent	$ for CPU
2012	14×3	4000	13	27000
2013	18×3	4000	17	27000
2014	23×3	4000	21	27000
2015	30×3	4000	28	27000
2016	46×3	4000	41	27000
Total in 5 years	131×3=393	20000	120	135000

Note. — The number of compute nodes purchased is based on the assumption that each node (26kSI2k, 104 HEP-SPEC 2006) would stay at the performance level of a node purchased in 2010. As the performance per node will increase over time the actual number of compute nodes after 5 years will be significantly smaller (probably O(70)), providing a combined performance of O(120) 2010 equivalent nodes. Note the factor of three redundancy for optimizing the distributed file system. Prices include 40% bulk discounts from purchasing through the RHIC ATLAS Computing Facility at BNL. **Power, cooling and maintence will be provided at no extra cost to this experiment.**

recently has been maintained by increasing the number of cores. The Jarvis and Sheldon code can make use of multiple cores and thus fully utilize high memory and multiple cores optimally. Future algorithms will be coded for multi-core performance also.

Note also, we plan to support shear algorithms from groups other than Jarvis & Sheldon. It is difficult to predict the speed of these algorithms, but for example first tests of the imcat based pipeline provided by Mandeep Gill is a factor of five slower than the Jarvis & Sheldon pipeline. Certainly this can be improved, but we may expect the multiplier for adding additional algorithms to be greater than unity.

We must also permanently store the ∼300 TB of data in order to efficiently process it through multiple algorithms.

5.2 Purchasing Plan

Taking the fiducial cluster of 80 nodes and the desired storage, we have developed a purchasing plan that should be nearly optimal in the sense that we can process data as it arrives but take advantage of increasing computing power and storage per dollar. This plan is outlined in table 2.

To store and process the ∼300 TB of DES data, we propose to spend about $30,000/year for five years. The first year we will acquire ∼ 13 nodes with at least 32G of memory each (26kSI2k, 104 HEP-SPEC 2006). In following years we will purchase more computing for the same price, and keep a trajectory to our goal of about 70 new nodes. Note, the table shows **2010 equivalent nodes**; adjusting for a compounded Moore's law, we get 122 of today's nodes corresponding to 70 actual nodes. These 70 nodes will augment the nodes we currently have, but note the existing nodes are in heavy use for other purposes.

Note in table 2 we have listed storage purchases per year. We are also requesting and additional 30TB of disk to provide storage for the lensing data products, including many re-processings. This makes a total of ≈ 130TB of storage, with a factor of three redundancy.

It is important to note these prices include bulk discounts of order 40%. This is due to purchasing along with other BNL experiments through the RHIC ATLAS Computing Facility (RACF). Because these resources are on a relatively small scale for the RACF, power, cooling, and maintenance are provided at no additional cost.

6 Brookhaven as a Host for the Computing Resources

Brookhaven is well suited to hosting this computing initiative.

Erin Sheldon of BNL is an expert at performing weak lensing analysis in enormous datasets. He has performed many lensing analyses using data from the SDSS, which is the largest lensing data set to date. He has been a member of DES since 2003 and has since helped to develop the current de facto pipeline used for lensing. He has also developed a general framework for processing single epoch and multi epoch DES data through any code. This framework will support various DES lensing algorithms.

BNL will support any DES weak lensing efforts as needed. Current development of the Jarvis/Sheldon pipeline is primarily occurring at BNL, and all major recent tests and runs

of the code have occurred there including DC6b. Catalogs from individual runs of the code are available on the BNL web site.

Mandeep Gill of Ohio State and Tim Eifler have already begun working at BNL and Mandeep will be the first to incorporate an alternate lensing codes into the framework. These catalogs will be hosted at BNL with a possible future release through the Brazil portal.

The Brookhaven RHIC ATLAS Computing Facility is massive and world class. In comparison to our plan for ∼70 new machines, the other experiments sharing the RACF support about 7500 equivalent cpus, many petabytes of storage, and three supercomputers. Our system will use power in the kilowatt range, whereas currently RACF uses 2.5 megawatts continuously. In preparation for the data coming from ATLAS, the computing center will more than double in size and power usage during the period we will purchase our computers. Because our needs are insignificant in comparison, they will provide us with complimentary power, cooling, and maintenance, as well as bulk purchasing discounts of order 40%. The RACF is an excellent base upon which to build our computing initiative.

References

[1] E. S. Sheldon et al. Cross-correlation Weak Lensing of SDSS Galaxy Clusters. III. Mass-to-Light Ratios. *ApJ*, 703:2232–2248, October 2009.