Computing Resources for DES Weak Lensing

Erin Sheldon
Brookhaven National Laboratory

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.

We request funding for computer hardware at $100,000/year for five years in order to 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 ∼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 has recently joined BNL as a postdoc and will work testing the WL pipeline and developing analysis codes. Mandeep Gill of Ohio State has begun working on an alternative WL code at BNL. Any other interested parties are encouraged to join this effort at no cost to them. Tom Throwe, Brookhaven computer systems and software 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 DESDM, but is scheduled to be run 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 better than a few percent. But the noise per galaxy shear estimate due to the intrinsic shape of the galaxy is of order 30%. 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 processing is complete, we must also analyze the data to extract parameters of the dark matter and dark energy. 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 "Bach" cluster at BNL. The Bach cluster has three compute nodes, each with 8 Intel Xeon 3GHz cores and 32GB RAM.

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 DC4 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 coadd images made use of a subset of these images, about half.

For SE images (individual 4k by 2k CCDs from a given pointing), there are number of steps in the 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

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

Data Set	# images/tiles	Size	Memory Per Job	CPU hours
DC4 SE	46,500	700G	1G	1728
DC4 ME	22,402/224	355G	15-48G	576

Note. — Resources used for processing i-band DC4 images using the Bach cluster at BNL. The Bach cluster consists 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 Bach machines did not have high memory.

associated pixels and PSF. For our current code, this takes about 2.5 minutes per image. The processing of each image is entirely independent.

For multi-epoch 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 previous processing described above. We then perform a joint fit for the shear across all images. This takes about twenty minutes per coadd tile using all 8 cores in parallel, and works out to be about 5 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 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.

5 Extrapolation and Requested Resources

The above processing timings were based on about 1Tb of i-band data. The full final DES data will be of order 1Pb. We can simply multiply the timings by 1000 to get a rough

Table 2. Projected Computing Purchases

Fiscal Year	Disk Storage [TB]	$ for Storage	Compute Servers 2010 Equivalent	$ for CPU
2010	130	55,000	15	45,000
2011	162	55,000	19	45,000
2012	203	55,000	23	45,000
2013	254	55,000	29	45,000
2014	318	55,000	36	45,000

Note. — The number of compute nodes purchased from 2011 on 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(122) 2010 equivalent nodes. 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.**

extrapolation. Thus for our current algorithms and current CPU speeds, it would take about 5 months to process the full DES data set on an 80 node cluster of equivalent machines. We fully expect to speed up the algorithms by a factor of a few. But ignoring that, we will also gain as processing power follows Moore's law. This gain has traditionally been in transistor density on a single device, but 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. Extrapolating these trends we should be able to process the full data set in two or three months. This would allow many re-processings for a few different weak lensing measurement algorithms, and facilitate science analysis of the results.

We must also store the 1 Pb of data in order to efficiently process it through multiple algorithms.

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 petabyte of DES data, we propose to spend about $100,000/year for five years, accruing about 200Tb of storage per year plus processing nodes. The first year we will acquire 130Tb of storage and 15 8-core nodes with 32G of memory each (26kSI2k, 104 HEP-SPEC 2006). In following years we will purchase more disk for the same price, and keep a trajectory to our one petabyte goal. We will assemble 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 3 nodes we currently have and the additional ten nodes BNL has recently purchased, which will come online in Fall 2009.

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 leading up to DC5. Catalogs from individual runs of the code are available on the BNL web site.

Mandeep Gill of Ohio State has already begun working at BNL and will be the first to incorporate his lensing codes into the framework. These catalogs will be hosted at BNL for the collaboration to test until ready for official release, at which time they may be released through the portal.

The Brookhaven RHIC ATLAS Computing Facility is massive and world class. In comparison to our plan for $\sim$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. *Accepted ApJ, appearing*, October 2009.