North American
ALMA Science
Center

ALMA data rates and archiving at the NAASC

NAASC Memo 110

Mark Lacy, David Halstead

Date: 2012 April 26

ABSTRACT

The ALMA baseline correlator can send 1 GB/s of lag data to the data processing cluster, which after processing and conversion to the ALMA data format would result in 512 MB/s of ALMA "raw" data (visibilities and autocorrelations). In practice, the data rate from ALMA is much lower due to a combination of scientific and practical considerations. The current ALMA operations plan is based on a data rate of 200TB/yr, an estimate obtained from consideration of specific science cases in the Design Reference Science Program. This study was, however performed some years ago. In this memo we attempt to update this estimate based on experience from ALMA Cycle 0 and developments in the scientific field since then. We suggest that, unless a clear policy to limit data rates is defined, that ARCs budget for up to 700TB/yr of archive storage and a data link speed to Chile of 100-300Mb/s for Full Science operations, ramping up to 1Gb/s mid-decade. We describe how the NAASC will archive this data volume, given the constraints of power and rack space in the Edgmont Road building.

1 Introduction

This document is structured as follows. The first section discusses the science cases used to estimate data rates for ALMA, and their possible future evolution. The second section describes the physical limitations on data transfer and how they might evolve on an ~10yr timescale. We then discuss the likely seasonal variation in the data rate before concluding with a discussion of data management strategies that could be employed.

2 Data rates for science projects

2.1 Early estimates of the ALMA data rate

ALMA memo 501 (Lucas et al. 2004) makes an estimate of the data rate for ALMA based on a selection of projects from the ALMA Design Reference Science Plan (DRSP). This study concluded that the mean data rate from ALMA would be 6MB/s, with a peak rate of 60MB/s. This estimate has increased slightly since then due to an ~10% addition from the compact array (ACA) correlator to 6.7/67MB/s. (The ACA correlator is capable of data rates up to 2GB/s, however, we believe it will be typically used in the same modes as the main array to complement the *uv*-coverage of 12m observations, thus the 10% estimate is reasonable.)

Assumptions were:

1. Images have a spatial sampling 1/3 the beam, and only final images are stored. (This leads to images taking up about 5% of the total data volume.)

2. 4 bytes/visibility

3. Only the part of the spectrum required by the observer to satisfy the DRSP science goal is kept, and it is sampled at the Nyquist frequency corresponding to the required resolution.

4. Only one (WVR corrected or not) dataset is archived

5. Integration (sampling) time $82/b$ where b is the maximum baseline in km, up to a maximum of 45s.

6. Calibration follows standard procedures.

2.2 Estimates based on ALMA Cycle 0

The NAASC has used the NA Cycle 0 programs to estimate the data rate based on extrapolation from a 16 element to a 50 element main array. We found that Cycle 0 programs differed from the DRSP programs in two important respects:

1. Multiple FDM basebands were the rule rather than the exception, and no on-line channel subsetting or averaging was available, thus a typical project had ~4000-16000 spectral channels instead of a few hundred. In many cases extra basebands were set to contain "bonus lines" that were not the main science goal of the proposal, but which would add value to the observations.

2. Integration times were not tuned to the baseline. Instead all FDM observations were obtained with 6s averaging and all TDM ones with 2s.

Item (1) above reflects the way mm/submm science has changed since the DRSP was written. Many more interstellar molecules are now known, and the wide bandwidth of ALMA means that there is a high probability of multiple lines being available in a single Science Goal setup.

Table 1 shows the results of assuming a Cycle 0 observing program in Full Science. The last three columns show that, based on this, data rates in FS >400TB/yr are to be expected, and it would be prudent to allow for data rates as high as 700TB/yr, and up to 1PB/yr when the AOS network upgrade is carried out. These numbers do allow for some buffer as they assume 100% observing efficiency, however we believe this is prudent in the light of the uncertainty in these figures.

MODE	FRACTION OF TIME REQUESTED IN CYCLE 0 NA SBS	FULL SCI DATA RATE 1S SAMPLING CAPPED AT 64MB/S (MB/S)	FULL SCI DATA RATE 1S SAMPLING CAP (MB/S)	FULL SCI DATA RATE (2S TDM 6S FDM SAMPLING) (MB/S)	OVERALL MEAN DATA RATE (ASSUMING 30% 1S SAMPLING, 70% 2S/6S SAMPLING CAPPED AT 64MB/S) (MB/S)	OVERALL MEAN DATA RATE (ASSUMING 30% 1S SAMPLING 70% 2S/6S SAMPLING NO CAP)(MB/S)
4TDM	0.37	4.5	4.5	2.3	2.93	2.93
1FDM	0.063	34.1	34.1	5.7	14.2	14.2
2FDM	0.025	64.0	68.1	11.4	27.2	28.4
3FDM	0.02	64.0	102.2	17.0	31.1	52.3
4FDM	0.52	64.0	136.2	22.7	35.1	56.7
Weighted mean data rate (MB/s)	-	40.0	78.4	13.6	**21.5**	33.1
Weighted mean Data rate (Mb/s)	-	320	627	109	**172**	264
Weighted mean data rate (PB/yr)	-	1.26	2.47	0.43	**0.68**	1.04

Table 1: The top rows show the data rates from each mode according to various assumptions. The bottom three rows (shaded) show the weighted mean data rates assuming the modes are distributed as for Cycle 0 NA proposals in various units (Mb/s for transfer rates, PB/yr for storage considerations, "most likely" estimate in bold). Full science data rates assume 30% of data taken in bands 9 or 10 and/or extended configuration, thus requiring 1s time sampling. The remainder have time sampling of 2s (TDM) or 6s (FDM) as in Cycle 0 SBs. The 64MB/s cap is imposed by the network at the AOS, an upgrade of this would remove the cap. Note that these estimates assume 100% observational efficiency, unlikely to be achieved in practice (though this ensures a 20-30% buffer in case these rates are underestimated).

2.3 Effect of future cycle capabilities

The extent to which Cycle 0 provides a good template for future cycles is debatable. Future capabilities may both increase and decrease the data rate:

1. *The ability to change the integration time.* Projects in Cycle 0 were observed with shorter integration (sampling) times than required by the $82/b$ law. The principal reason for this was that the WVR corrections needed to be applied offline. When online WVR correction becomes available these may be increased. However, for projects requiring high dynamic range, short integration times (~1s) will still be preferred to allow effective self-calibration of the data. Thus it is unclear how the ability to tune integration times will affect the data rates in the future. For the purposes of this document we assume making the integration times variable will have a net zero impact on the overall data rate.

2. *Facilities to subset or average the channels.* These will result in a decrease in the data rate, up to a factor of 10 for some projects (e.g. many extragalactic line detection and gas dynamics experiments that are not already using the TDM correlator mode - perhaps 10% of projects overall). This fraction could be increased if proposers were prepared to forgo serendipitous lines in their projects, for example, if there were a penalty for high data rate proposals.

3. *On-the-fly (OTF) mosaics*. As discussed below, this is an example of an observing mode with a strong science case for using high data rates, which will be commissioned in a future Cycle (2+). Assuming ~10% of proposals are OTF and require high data rates this will likely cancel out any gains from item (2), averaging or subseting the channels, to first order.

2.4 Science cases for future data rates

At present, science cases can easily be written (and were, in Cycle 0) for 4FDM full resolution (3840 channels) basebands. If self-calibration is possible, or if using very extended arrays and/or high frequencies, time resolution down to 1s may be required. Such observations would already exceed the current 64MB/s data rate cap on the correlator output by over a factor of two, placing us into Gb/s territory and requiring a 10Gb/s upgrade to the AOS network. On the other hand, science cases for extremely high data rate (~10Gb/s) projects will be limited. Only a few observers will be interested in high (sub km/s) resolution over the entire 8GHz ALMA bandwidth at high frequency and/or in extended configurations, and an 8hr dataset taken using such data rates would amount to ~30TB, which brings with it considerable processing challenges.

On-the-fly (OTF) interferometry provides a specific example of a strong science case that will lead to a severe data rate challenge. OTF modes will require short sampling times, down to the limits of the correlator modes (512ms would allow all FDM correlator modes, 16ms for TDM). Data rates as high as ~300MB/s could be justified scientifically fairly easily, for example in the case of a line survey of a bright starforming region. With the current 64MB/s limit, OTF modes would have to be restricted in either number of channels, scan speed, or both.

Data reuse is also becoming increasingly important in astronomy, even for telescopes such as the Hubble Space Telescope that, like ALMA, are not primarily survey instruments. The archival value of taking large data volumes should therefore not be underestimated. In the future, upgrades to the correlator, receivers and electronics on an ~10yr timescale that allow a significant increase in bandwidth beyond the current 8GHz may be considered. This would considerably expand the science that could be done with the array, and lead to further good science cases for very high data rates. Such an upgrade would put even a 1Gb/s link under pressure during high spikes (Figure 1).

The question of whether duplicate ASDMs will need to be kept while the online WVR calibration is being assessed is still open, this will of course double the effective data rate during that assessment period, and lead to short-term pressure on the archiving and data transfer.

3 Data rate limitations

3.1 Correlator/data capture

The ALMA correlator can output 1GB/s (Pisano et al. 2005) (though this assumes 8-byte visibilites, which are the size output by the correlator, in practice, however, 4-byte visibilities are typically archived, making the effective maximum data rate 512MB/s), this is currently limited to 64MB/s due to the speeds of the network interface cards and the 1Gb/s connections used. An upgrade would be possible at relatively modest cost to upgrade the connections and network to 10Gb/s allowing the full potential of the correlator to be realized, and is likely to be proposed as part of the ALMA development plan.

3.2 Transfer

AOS to OSF:

Data is transfered from the AOS to the OSF over a fiber link with a current capacity of 1Gb/s, though there is a plan to upgrade to a 10Gb/s link.

OSF to SCO:

Data is currently transferred to the mining town of Calama, between San Pedro and the coastal city of Antofagasta via a microwave link, currently limited to 100Mb/s. At Calama it joins the Chilean fiber backbone for transfer to Santiago. 20% of the 100Mb/s is reserved for VOIP telephone system. The JAO would like to upgrade this to a minimum of 300Mb/s (possibly as high as 2.5Gb/s) on a 1-2 year timescale, either with further microwave links, or an embedded fibre. The fiber is the preferred solution.

JAO internal documents suggest that a target of 500Mb/s is being proposed on a 2012 timescale, with an OSF to Calama fiber, commercial leasing from Calama to Antofagasta then REUNA from Antofagasta to SCO. Further upgrades to 2.5Gb/s have been proposed on a 2013 timescale.

SCO to the ARCs:

NA has secured a 100Mb/s link from SCO to Florida International University, Miami and hence to the US research backbone (I2/NLR). Cost is $50k/year, negotiated as a share of a 622Mb/s link used by AURA/NOAO-CTIO. The upgrade path would see this AURA link increased to a 1Gb/s link in the near future, and a 10Gb/s link mid-decade to support NOAO initiatives such as LSST, with NRAO retaining a minority share. Our plan is to negotiate an increase in bandwidth up to 300Mb/s at the start of Full Science (end of 2013), and increase mid-decade to a full 1Gb/s (3.9PB/year) to cover the anticipated increase in data rate from an AOS network upgrade. All these links are planned to be burstable to capacity, to allow the full bandwidth to be used, for example, to take advantage of the fact that most optical telescope data transfer will take place at night, leaving the daytime free for ALMA.

EU has an agreement through ESO with REUNA for a 20Mb/s link, this may be upgradable in the near future, however. In addition they have available any unused portion of the ESO 30Mb/s link (this link is heavily used at night, but less so during the day). The link to EA is currently 10Mb/s, with a further 10Mb/s on a "best efforts" basis, this will be upgraded to 25Mb/s with a further 25Mb/s "best efforts". The long-term plan for EA is still TBD.

3.3 Data storage

The NAASC has 500PB/year of storage in its budget to 2015, corresponding to a steady-state data rate of 100Mb/s. Upgrades beyond 1PB/year would mean outsourcing the archive to a computing center in about 2014 (earlier if EVLA and GBT archives need to be held on spinning disk in CV) as the computer room cooling capacity will be exceeded. Other ARCs are building to the 200TB/yr in the Ops Plan D.

4 Seasonal/Cyclical variations

ALMA will work on a 1 year cycle, with configurations varying from compact through extended. Extended array observations will use higher data rates as the data need to be sampled more often. Extended array observations also typically require better conditions, as phase coherence needs to be maintained over longer baselines. High frequency (>500GHz) observations are also likely to require high sampling rates, and will also be concentrated towards times of year when the conditions are good. Strong seasonal variations in data rate are thus to be expected. http://almascience.eso.org/about-alma/weather shows the monthly variations in water vapor at the ALMA site. On this basis the best months (for extended arrays and/or high frequencies, and thus high data rates) will be July through October, the worst months (compact arrays, low frequencies, or no data at all) will be January through March.

Table 1 assumes a 30:70 split for high time sampling versus "normal" projects. The data rate during these periods is likely to be about three times higher than average. Figure 1 shows a notional estimate of the growth of the ALMA data rate with time out to ~2020. The seasonal variations assume that the winter quarter contains most of the high frequency/extended configuration observations, leading to spikes in the data rate at those times of year. Until we are able to furnish a 1Gb/s link to Chile (2015/2016) these spikes will be difficult to manage with network transfer alone.

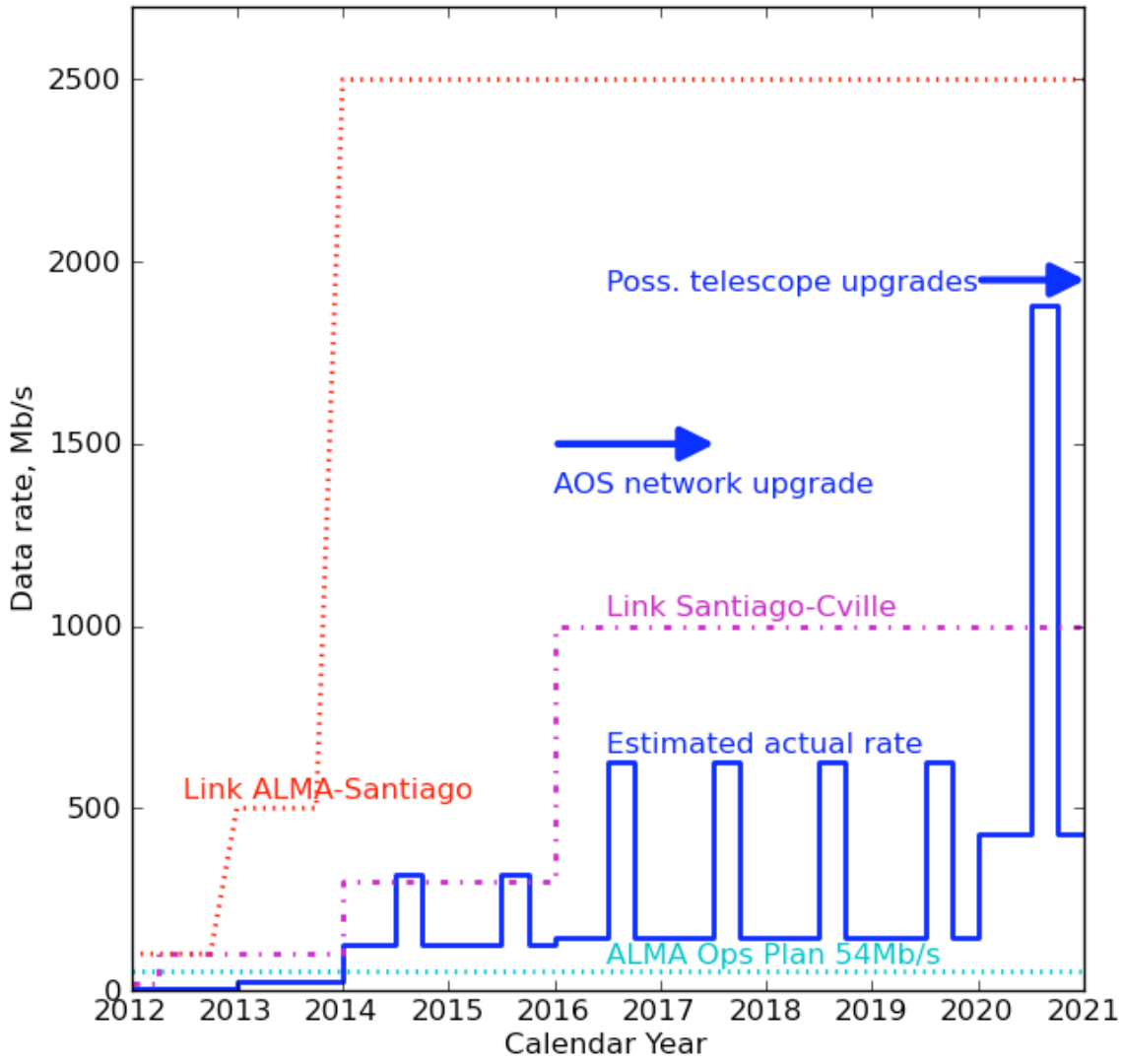

Figure 1: likely growth of the ALMA data rate through 2020. The blue solid line is our best estimate of the data rate based on Cycle 0 proposals and likely seasonal variations, assuming an AOS network upgrade in 2016 and a possible observational bandwidth increase on the 2020 timescale ("Poss. telescope upgrades"). The dotted cyan line corresponds to the rate in the ALMA Operations Plan vD (200TB/yr). The red dotted line is the likely JAO target for the ALMA to Santiago link likely to be in place by the end of 2012. The dot-dashed magenta line is the NAASC link to Chile (100Mb/s from ~April 2012, with an upgrade to 1Gb/s mid-decade).

5 Growth of storage requirements in Charlottesville

The NGAS system consists of sets of 24-disk nodes, generally installed in sets of four. The first set of four nodes had 2TB disks, allowing for redundancy these result in 30GB of storage per node. Future nodes will use 3TB disks, or larger as they become available. Each NGAS node takes up 5Us of rack

space (in 40U racks) and consumes approximately 0.42kW of power. Power requirements in the ER computing room are dominated by the NAASC compute cluster (and the accompanying Lustre filesystem), each compute node occupies 1U of rack space and consume 0.27kW each. See Table 2 for the NAASC storage and compute budget.

In addition to the ALMA mirror, Charlottesville also needs to store the EVLA mirror and (selected) GBT data. Estimates are that each archive will reach ~1PB by the end of 2013. Limited power and cooling in the Charlottesville ER computing room means that we will need to outsource some storage in 2014+. UVa has recently opened a computing center where rack space is currently available for nominal cost to collaborations. We will therefore use our collaborations with UVa to try to secure some of this space for data which is public, i.e. past its proprietary period. Alternatively, should negotiations with UVa prove unsuccessful, we will ask NCSA to host data for us. As all our archives are mirrors, and we will only outsource storage of public data, there is no data security risk in this.

6 Implications for data management strategy

6.1 Managing the data rate growth through to Full Science (2013)

A reassessment of data rate policy is urgently required, otherwise data rates in Cycle 2 will exceed those in the Operations Plan by at least a factor of two. There are two options - reduce data rates (at the expense of lost science) or increase the budgets for data transfer and storage. Data rates can be reduced by penalizing high data rate proposals (see section 5.5). Time sampling could be more carefully examined to ensure that no more data was taken than needed in a given configuration. As more correlator modes are commissioned, the flexibility to target relatively narrow spectral regions at Nyquist sampling becomes available, this could actually lead to a decrease in the data rate if there was some penalty for high data rate proposals. The decision as to whether to use the online WVR correction or not could be made during a limited campaign on SV datasets rather than duplicating corrected and uncorrected ASDMs in the archive for all science observations.

The alternative to reducing the data rates is to pay for high data rates, this could be done provided the infrastructure was in place. For example doubling the current 100Mb/s data rate to Charlottesville to 200Mb/s might cost ~$50-100k/yr, plus ~$50k for storage of every extra 100TB in the archive. If internet bandwidth to South America is the fundamental limitation, disk shipping from JAO to the ARCs remains an option, although this would require some software effort (as NGAS would need to be able to utilize both disk and internet transmission of bulk data, possibly at the same time). Disk shipping would also require more Data Analysts at both the JAO and the ARCs to perform the disk copying. Estimated cost for this would be 2DAs/ARC (one at the ARC, one at JAO) plus the cost of disks and shipping, ~$200k/yr per ARC, $600k/yr in total for the project. Another possibility if the link from Santiago to the ARCs is a limiting factor would be to transfer the data to a single ARC that has a good link to Chile, and then distribute from that ARC to the other two using peer-to-peer or similar technologies. Again, this would require software development.

Early science data processing also results in a data rate challenge. Unlike the case in Full Science, when the only large files to be archived will be the ASDMs, calibration and flagging tables, and final images, we are currently (March 2012) archiving in addition two measurement sets, the ASDMs with the WVR and Tsys calibration tables applied, and the final calibrated

measurement set. Each of these can be up to twice the size of the ASDM (less if time averaging is applied). Towards the end of Cycle 0, when there are about 30 antennas and the array is observing for ~60% of the time, this will result in a data rate comparable to that in Full Science. We anticipate this situation may be untenable unless the ARCs and JAO are able to increase their bandwidth ahead of schedule, or ship disks to maintain the mirror archives. Therefore we are likely to change the packages and ship either one or neither of the measurement sets.

6.2 Managing data rate growth 2013-2020

In the longer term, the costs of the high data rates available to ALMA need to be weighed against their benefits. Costs for both transfer and storage are likely to decrease with time, but at an unknown rate. For the NAASC, the availability of a 1Gb/s (3PB/yr) link on a ~2015-2018 timescale will remove data transfer concerns for all but the most extreme datasets. (The availability of such a link may depend upon the scheduling of LSST construction, however, which is yet to start construction.) These data will still need to be stored, however, and storage may need to be outsourced to University of Virginia or the US National Supercomputing facilities, where economies of scale will allow cheaper storage (possibly as low as $200-300/TB, compared to $500/TB for local storage).

6.3 Beyond 2020

Plans to increase the bandwidth to ~20-100GHz so as to be able to take a spectrum of an entire ALMA band in one shot may not be totally out of scope for development proposals, leading to an archived volume ~10PB/yr. A new correlator to accompany this could also be on the horizon on the 10-20 year timescale as software improves.

6.4 Other impacted software systems

Besides the archive, higher data rates will have impacts on other sub-systems. Data processing clusters may need to be increased in size, and software parallelized further to keep processing speeds high enough to avoid backlog. Similarly, online programs run at the OSF such as TelCal, quicklook pipeline and QA0 and QA1 software, may need to have upgraded hardware and software to deal with higher data rates.

6.5 Limiting data rates on proposal submission

In general, it would be best not to limit the possible science with ALMA due to data rate considerations given the relatively low cost of data management as a fraction of the total observatory cost. However, there may be "crunch points" that arise when the data rate will exceed our capability to transport, process or store it, and we need to limit the data rates of accepted proposals. A practical way to do this without completely disallowing the science from high data rate proposals would be to place projects requiring significantly more than the target average data rate (say 12MB/s at present) into a restricted pool of proposals with a small number of hours (say 10% of the available time) allocated, resulting in them needing to pass a higher bar to be scheduled on the telescope.

References:

Pisano, J., Amestica, R. & Perez, J. 2005 10th ICALEPCS Int. Conf. on Accelerator & Large Expt. Physics Control Systems. Geneva, 10 - 14 Oct 2005, PO2.067-4

NAASC HPC hardware and facilities running costs

Cost per drive	$300
Cost per NGAS/Lustre node	$12,200
kW Power per NGAS (VA/0.83PF)	0.48
Cost per compute (with network)	$4,000
kW Power per cluster node	0.42

START REFRESH

FY=	2011	2012	2013	2014	2015	2016
ALMA products (TB)	2	20	650	680	680	680
Total Data (TB)	2	22	672	1352	2032	2712
Drive Size (TB projected beyond 2(	2	3	4	5	6	7
New NGAS node count	4	4	8	8	8	8
Total NGAS node count	4	8	16	24	32	36
New NGAS Storage (TB)	120	180	480	600	720	840
Total NGAS Storage (TB)	120	300	780	1380	2100	2820
New Lustre node count	2	2	2	2	2	2
Total Luster node count	2	4	6	8	10	10
New Lustre Storage (TB)	60	90	120	150	180	210
Total Lustre Storage (TB)	60	150	270	420	600	750
New Compute node count	8	24	16	16	0	8
Total Compute node count	8	32	48	64	64	72
Power needed (kW)	6.2	19.1	30.6	42.1	47.0	52.2
Power cost ($ per kWh)	$0.11	$0.12	$0.13	$0.14	$0.15	$0.16
Power cost per year (facilities $)	$7,515	$25,142	$43,618	$64,615	$77,201	$91,578
New Arch/HPC System cost	$105,200	$169,200	$186,000	$186,000	$122,000	$154,000
Out-year maintenance (10%)	$0	$0	$0	$10,520	$27,440	$35,520
Annual HPC/Archive cost	$112,715	$194,342	$229,618	$261,135	$226,641	$281,098

Excludes licenses, network, shipping and non-HPC/Archive hardware costs (~$230k/year by 2013