1. ALMA Pipeline Cluster specification

The following document describes the recommended hardware for the Chilean based cluster for the ALMA pipeline and local post processing to support early science and commissioning. The hardware specification closely mimics the proposed EVLA post-processing cluster at the NRAO/NM DSOC and reflects current understanding of the current state of CASA, expected improvements and final parallel implementation.

The document describes specific hardware for three distinct subsystems: cluster processing nodes, high speed I/O data store and high speed network to connect the two to each other and the local NGAS based archive described elsewhere. The document is only a physical description suitable for ordering and physical installation. Configuration and implementation will be covered elsewhere.

The resulting cluster will consist of 8 processing nodes connected via 40Gbit QDR Infiniband to a Lustre parallel filesystem. The proposed nodes can process ~400MB/s within the major cycle of the CASA imaging task. The proposed filesystem will provide ~60TBytes of addressable storage and can supply individual nodes with data at a peak rate of ~750MB/s per node and an aggregate rate of 3GB/s for the entire cluster.

Lastly, most development to date regarding CASA parallelization and HPC specifications has focused on continuum imaging with the EVLA. We believe we understand how CPU, I/O and memory demands will translate to spectral line imaging with ALMA but they are ultimately extrapolations and subject to change. To that end the proposed system is designed to be adaptable, particularly where memory requirements are concerned.

2. Compute processing node specification: $26K

The initial ALMA Pipeline cluster will consist of 8 compute nodes. These can be standard Dell 410 model 1U servers.

Dual 2.4Ghz six core E5645 Nehalem processors provide the best price performance and most balanced MByte/s processing with respect to the backend Lustre filesystem's ability to provide data to the processing nodes.

The Dell 410 model server nodes have 8 total DIMM slots. Initially the nodes should be purchased with 24GB of memory 3 x 4GB DIMMs per processor. The Nehalem E5645 processors have 3 bus lines into memory. Filling the 4th DIMM slot should only be done if the memory is needed as there is a slight performance penalty.

There is no local storage on the processing nodes, only a local OS disk is required. Remote Lustre storage is accessed through a 40Gbit QDR Mellanox MT26428 PCI-E 8x HCAs. An example specification is included below in Figure 1. Each node should cost ~$3125; total node cost is ~$25K.

3. Distributed Lustre filesystem specification: $25K

Lustre, a distributed parallel filesytem, will be used to store datasets for processing. Lustre presents each processing node with the same filespace so there is no need to redistribute data from a central filesystem to the individual nodes.

A Lustre filesystem consists of a Metadata Server (MDS) which manages file layout and tracks metadata changes (i.e. ownership, permissions, size, etc) on its Metadata Target (MDT) and one or more Object Storage Servers (OSS) which stores actual disk blocks. Each OSS contains one or more Object Storage Targets (OST) which are discrete physical storage devices, typically RAID arrays.

For the proposed system the MDS is a Dell 410 1U rack mounted server configured the same as a processing node with the following differences. The MDS has two 250GB internal disks in a software RAID-1 mirror to store the OS and MDT. The MDS only requires 4 to 8GB of memory not 24GB.

There are two OSSes, each is a 4U 24disk chassis with a Superlogics X8DTH-i motherboard with 7 PCI-E 8X slots, dual E5520 Xeon processors, 4GB of memory and redundant hot swappable 1200 watt power supplies. In addition each OSS supports four OSTs. Each OST consists of an eight port 3ware 9650SE raid controller attached to one of the PCI-E 8x slots and 6 Western Digital WD2003FYYS 2TB hard drives.

Each OST is ~7.5TB of addressable storage and can sustain 400MB/s sequential reads or writes. The 4 OSTs in an OSS have an aggregate volume of 30TB and can sustain 1.5 to 1.6GB/s sequential reads or writes.

The MDS and each OSS are connected to an Infiniband switch via the same 40Gbit QDR Mellanox MT26428 PCI-E 8x HCAs as are in the processing nodes.

A drawing showing the layout of each OSS and their connection to the switch and processing nodes is provided below as Figure 2. Each OSS costs approximately $11K and the MDS costs a further $3K; total cost for the Lustre filesystem is ~$25K. A parts list description of an OSS node is included in Figure 3.

4. High Speed Infiniband Network: ~$10K

A small 40Gbit QDR Infiniband switched fabric is used to connect processing nodes to the Lustre filesystem and optionally to the local NGAS archive. While both 10Gbit Ethernet and 40Gbit QDR Infiniband provide sufficient bandwidth to individual nodes Inifiniband is lower latency than 10Gbit Ethernet and less expensive over short distances in addition the Lustre OSSes can saturate a 10gbit Ethernet link.

Under normal circumstances Infiniband is limited to 15 meters. If the cluster and Lustre filesystems can't be in relatively close proximity to each other or the NGAS archive a 10Gbit Ethernet alternative could be considered.

The Inifinband fabric consists of a 36 port Mellanox IS5030 QDR switch plus a Fabric Subnet management license, the individual HCAs on each node and twinax cables of various lengths to connect each node to the switch. Cables part numbers are MCC4Q26C-00X where X is the length in meters; 4 meters are used as an example.

The following table lists items and costs for the Infiniband network

Description	QTY	Unit cost	Total Cost
MIS5030Q-1SFCMI switch	1	5100	5100
MHQH19B-XTRMHQ HCA for processing nodes	8	415	3320
MHQH19B-XTRMHQ HCA for Lustre nodes	3	415	1245
MCC4Q26C-004 Twinax cabling	11	50	550

5. Total parts and costs: ~$61K

Total cost of processing nodes, Lustre filesystem and Infiniband network is approximately $60K. The total cost does not include infrastructure items like racks, power, PDUs, air conditioning etc.

Description	QTY	Unit cost	Total Cost
Processing Nodes	8	3250	26000
Lustre MDS	1	3000	3000
Lustre OSS	2	11000	22000
Infiniband Switch	1	5100	5100
Infiniband cards/cables	11	465	5115
Total			61,215

Dell PowerEdge R410

Price **$3,252.60**

Preliminary Ship Date: **4/29/2011**

My Selections | All Options

- **Dell PowerEdge R410**

Date	4/20/2011 10:35:40 AM Central Standard Time
Catalog Number	16 Retail rc986192

Catalog Number / Description	Product Code	Qty	SKU	Id
PowerEdge R410: PowerEdge R410 Chassis w/up to 4 Cabled HDs, Quad-Pack LED Diagnostics	R410CWG	1	[224-8691]	1
Ship Group: Shipping Material,PowerEdge R410	SHIPGRP	1	[330-4137]	2
Processor: Intel Xeon E5645 2.40GHz, 12M Cache, 5.86 GT/s QPI, 6C	E5645	1	[317-6154]	6
Additional Processor: Intel Xeon E5645 2.40GHz, 12M Cache, 5.86 GT/s QPI, 6C	2E5645	1	[317-1276] [317-6163]	7
Memory: 24GB Memory (6x4GB), 1333MHz Dual Ranked RDIMMs for 2 Procs, Optimized	24GR2PO	1	[317-5877]	3
Operating System: No Operating System	NOOS	1	[420-6320]	11
Internal Controller: SAS 6/iR SAS internal RAID adapter, PCI-Express for Cabled Configuratiion	SIRAPCI	1	[330-7658] [342-0767]	9
Hard Drive Configuration: No RAID for SAS 6/iR Controller	NRH2C	1	[342-2701]	27
Hard Drives (Multi-Select): 250GB 7.2K RPM SATA 3.5" Cabled Hard Drive	250S35C	1	[341-9208]	1209
Power Supply: Power Supply, Redundant, 500W	500RNDT	1	[330-4141]	36
1st Hard Drive: HD Multi-Select	HDMULTI	1	[341-4158]	8
Embedded Management: Baseboard Management Controller	BMC	1	[313-7919]	14
Rails: ReadyRails™ Sliding Rails without Cable Management Arm	RRNOCMA	1	[330-4139]	28
Bezel: No Bezel	NOBEZEL	1	[313-0869]	17
Internal Optical Drive: DVD-ROM Drive, Internal	DVD	1	[313-7834] [313-9126]	16
System Documentation: Electronic System Doc, OpenManage DVD Kit with Dell Management Console	EDOCSD	1	[330-4148] [330-5280]	21
Power Cords: No Additional Power Cord	NOPWRCD	1	[310-9057]	38
Hardware Support Services: 3Yr Basic Hardware Warranty Repair: 5x10 HW-Only, 5x10 NBD Onsite	U3OS	1	[993-7452] [994-2610] [994-4019] [994-6019] [994-6058] [994-6627]	29
Installation Services: No Installation	NOINSTL	1	[900-9997]	32

Print

Figure 1 ALMA processing node

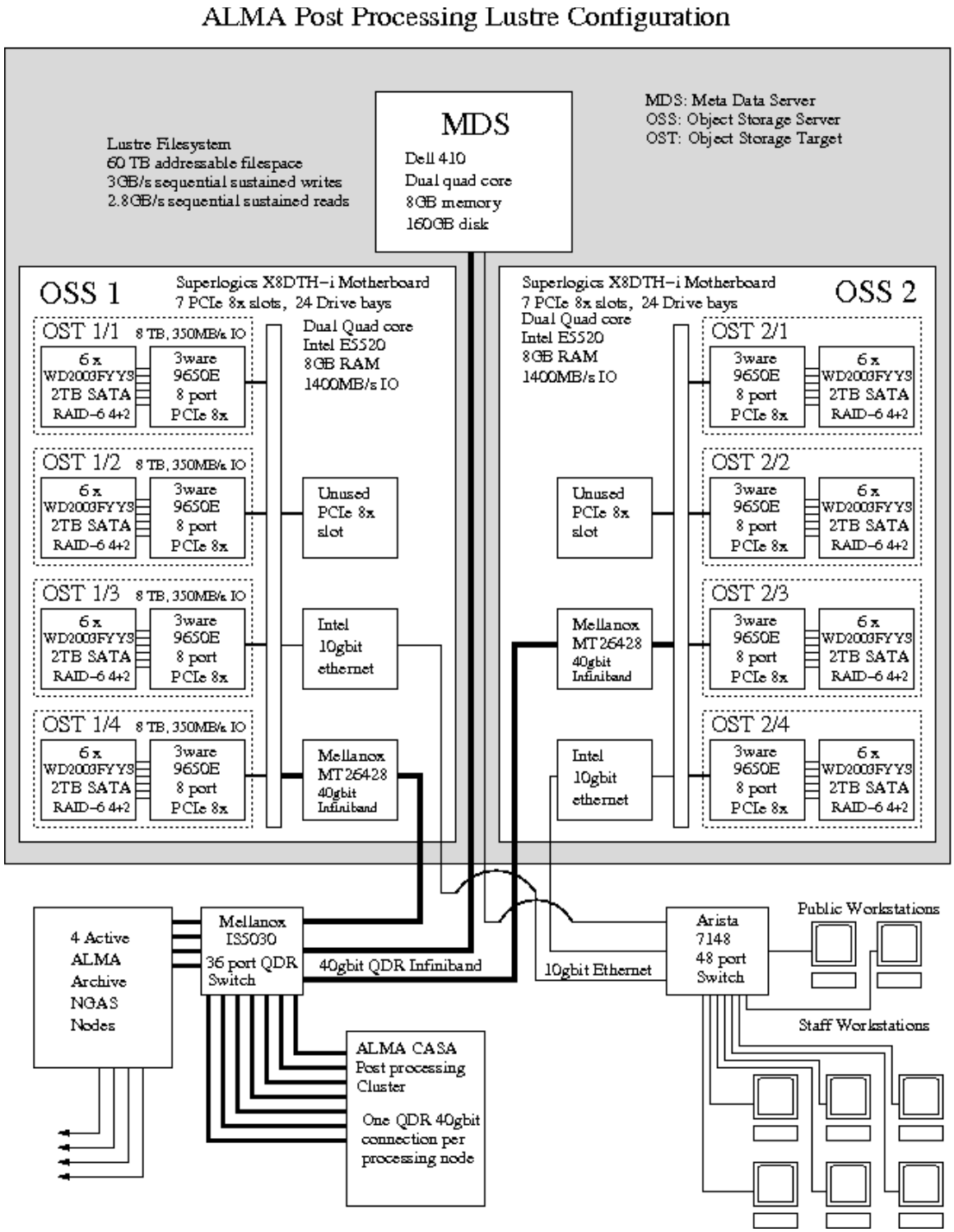


Figure 2 Lustre Diagram

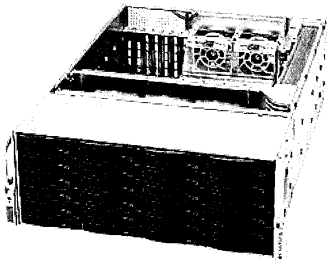
Detail Specification:	
Chassis:	4U Chassis w / 24 SAS/SATA hotswap bays , 1200W high-efficiency (1+1) redundant power supply
	
CPU:	Dual Intel Xeon E5520 / 2.26 GHz - LGA1366 Socket - L3 8 MB
Memory:	4GB DDR3 1333 Memory
MotherBoard:	SUPERMICRO X8DTH-i - Motherboard - extended ATX - Intel 5520 - LGA1366 Socket – SATA-300 (RAID) – video with 7 PCI-E 2.0 8x slots
Video:	Built in Video
Hard drive:	24 - Western Digital 2TB WD2003FYYS 64MB cache 1- 80GB OS disk
LAN:	Built in Dual Gigabit (10/100/1000) Ethernet Ports
Power Supply:	1200W high-efficiency (1+1) redundant power supply
Software:	Preload, OS installed for QA testing support Redhat ES4.2 or Linux 2.6.x kernel
Raid Controller:	[redacted] Four 8 port 3ware 9650 PCI-E 8x cards
CDDrive:	24X CD/DVD Drive
Rails:	Sliding Rail Kit included
Warranty:	Three Years – Parts and labor Warranty

Figure 3 Lustre OSS specification