Mem. S.A.It. Vol. 74, 415

Memorie della

Development of BEOWULF computer clusters for high performance parallel processing of astronomical images

G. Sedmak[1,2], E. Cascone[3], G. Russo[4], and S. Zaggia[5]

[1] Dipartimento di Astronomia, Università di Trieste, Trieste, Italy
[2] Istituto Nazionale di Astrofisica, Roma, Italy
[3] INAF, Osservatorio Astronomico di Capodimonte, Napoli, Italy
[4] Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Napoli, Italy
[5] INAF, Osservatorio Astronomico di Trieste, Trieste, Italy

Abstract. The project is aimed to the development of BEOWULF computer clusters for high performance parallel processing of astronomical images. The objectives, planning, and expected results are presented with reference to the 24 months schedule and the 200 keuro budget available on Italian MIUR COFIN 2001 national and local funds. The priority goal is to provide within 2003 the research units of Naples and Trieste with the medium-sized BEOWULF clusters and the application software packages needed to support and safely operate the pipeline processing for VST OmagaCAM and ESO cameras and spectrographs.

Key words. methods: numerical – techniques: image processing

1. Introduction

High performance digital processing of mosaic images is a task of increasing relevance to wide field and high spatial resolution astronomy, The development of telescopes such as the VST (Mancini et al. 2000) and VISTA (VISTA web page[1], 2001), equipped with imaging cameras of FOV of $1^\circ \times 1^\circ$ corresponding to 16k×16k pixels, led to realize two-dimensional CCD detectors larger than the maximum size of a single silicon chip. Case realizations are the European Southern Observatory (ESO) WFI camera (ESO WFI web page[2], 2001) and the OmegaCAM camera[3]. In these cases the format requested is realized as a mosaic of a number of smaller CCD detectors integrated into a unique support. The images output from a mosaic detector consist of an array of arrays of data. The different physical performances of the individual detectors enforce to calibrate separately the sub-images constituting the mosaic. This is needed in order to reconstruct the full

Send offprint requests to: G. Sedmak
Correspondence to: Dipartimento di Astronomia, Università di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy

[1] `http://www.vista.ac.uk/`

[2] `http://www.eso.org/projects/odt/WFI/imgal.html`

[3] `http://www.astro.rug.nl/~omegacam`

sized global image to make it ready for the following scientific data analysis. The mosaic calibration is done by means of appropriate dithering and stacking the mosaic sub-images in order to restore the gaps between the individual detectors of the physical mosaic. Telescopes and imaging cameras operated at high spatial resolution led to develop special procedures for the deconvolution of the observed Point Spread Function (PSF). The cases of space variant, poorly determined or unknown PSF are of particular relevance (Jeffries & Christou 1993; Boden et al. 1996). These applications include the restoration of sources convolved with interferometric-like PSFs and with the PSF of the atmospheric turbulence (speckle astronomy, see Roggemann & Welsh 1996), possibly extended to the tomographic restoration of wide field images observed with adaptive optics telescopes (Ragazzoni et al. 2000). The procedures needed can be generally operated in parallel mode through partitioning the images in arrays (mosaics) of conveniently smaller sub-images. The size of every subimage is determined in order to approximate the invariance of the PSF over the field with respect to the procedure used for the deconvolution. The processing and analysis of mosaic images request an appropriate amount of RAM memory and computing power as well as the mass storage for the images data base for the archive of the original (input) and processed (output) images. The configuration of the image processing system must take into account for the format of the mosaics, the number of mosaics, the number of sub-images in a mosaic and the images data rate in terms of number of mosaics to process and archive per unit time. Mosaic-oriented image processing procedures have been realized in astronomical standard software environments such as the NOAO IRAF (Valdes 1998; Valdes & Tody 1998) and ESO MIDAS (ESO MIDAS web page[4], 1999).

These procedures are compatible to parallel processing. The handling and processing of the data output from a camera of 16k$\times$16k pixel format (100 GigaByte every 24 hours) can be supported effectively by remote access supercomputing centers or by specially built local processing systems. The second option may show a substantially better performance to cost ratio. The mosaic structure of the images makes particularly convenient to use a gross grain parallel configuration with (at least) one processor for each sub-image of the mosaic. The computing and data base resources requested for the effective handling and processing of high spatial resolution images are quite large either for mosaics (deconvolution of large sized images with space variant PSF) and for series of images observed at high time resolution (speckle astronomy datacubes at a rate of 10 - 100 medium sized images per second). Also these cases can be conveniently supported by parallel computers but extra care must be given to the need of procurement and further development of suitable numerical algorithms and software codes. Prototypes of parallel computers based on commodity commercial components connected as nodes in a standard local area network have been studied (Warren et al. 1997; Reschke et al. 1996) and realized with positive results in the United States (Caltech Center for Advanced Computing Research web page[5], 2001) and in Europe (University of Heidelberg and University of Mannheim web page[6], 2001) using a number of Pentium class processors connected together through an Ethernet LAN and operated in Linux software platform. These studies led to define the parallel configuration BEOWULF (Ridge et al. 1997; Sterling et al. 1996) which shows very good performance and performance to cost ratio (5 GigaFLOPS per 50,000 US dollars

[4] `http://www.eso.org/projects/esomidas/doc/user/98NOV/volb/node61.html`

[5] `http://www.cacr.caltech.edu/resources/compute-resources.html`

[6] `http://suparum.rz.uni-mannheim.de/Comyc`

1998, see Sterling et al. 1998). In Europe two 8 nodes BEOWULF systems for astronomical applications have been built and are now under evaluation at the ESO (ESO BEOWULF web page[7], 2001). Up to now the Italian astronomical community considered but did not developed systematically high performance computers for mosaic image processing and in general for processing experimental data structured for parallel configurations. This is particularly relevant for the processing support of the observations expected from the new large telescopes and instruments. This recommended for the study and realization of suitable hardware and software systems for the Italian astronomy. The BEOWULF parallel approach can be a positive starting issue for the development of low cost, highly performant processing facilities. Following these guidelines a group of 2 research units of the Italian National Institute for Astrophysics (INAF) (Capodimonte Astronomical Observatory and Trieste Astronomical Observatory) and 2 research units of Italian University (Naples University Federico II and Trieste University) proposed a national research project for the development of BEOWULF computer clusters for high performance parallel processing of astronomical images. The project was positively refereed, approved and funded on the national COFIN 2001 fund of the Italian Ministry for Education and Research (MIUR) as a 24 months program with 200 keuro budget with 70% share covered by MIUR. The objectives, planned developments and expected results of the BEOWULF project are described in the following sections together with some preliminary results.

2. Objectives

The main goals of the program ordered by scientific and technological priorities are the following.

[7] `http://www.eso.org/projects/dfs/beowulf.html`

Technological goals:

(a) realise a BEOWULF cluster with 32 nodes specialised for gross grain parallel processing of mosaic images and release the system within 2003 as a national facility for the wide field images expected in particular from the VST OmegaCAM survey telescope;
(b) import in the Italian astronomical community the know how available at international level on gross grain parallel computing implemented by BEOWULF clusters of low cost commodity commercial components with special focus on handling and processing large images and high volumes of experimental data;
(c) test new protocols for fast network computing and optimise the impact of the latency times on global performance of BEOWULF clusters;
(d) test the usability and safety aspects of the BEOWULF clusters within a geographical network looking forward to the future availability of the fast Italian network GARR-C.

Scientic goals:

(e) development and test of BEOWULF application software for parallel processing of mosaic images of 8k×8k pixel format (ESO WFI) and 16k×16k pixel format (VST OmegaCAM);
(f) development and test of BEOWULF application software for parallel processing of high spatial resolution images characterised by space-variant morphological properties.

3. Research group

The national research group includes 4 research units as follows: (a) Capodimonte Astronomical Observatory, local coordinator E. Cascone. (b) Naples Federico II University, local coordinator G. Russo. (c) Trieste Astronomical Observatory, local coordinator P. Santin, executive coordinator S. Zaggia. (d) Trieste University, local coordinator G. Sedmak. The national coordinator is G. Sedmak.

Computer	Num Procs	Rmax(GFlops)	Nmax(order)	Rpeak(GFlops)
IBM SP2 (160 MHz)	64	29.45	27500	41
CRAY T3E-1200 (600 MHz)	32	25.98	42240	38
Cray T3D 256 (150 MHz)	256	25.3	40960	38
Sun Ultra HPC 10000(250 MHz)	52	21.68	19968	26
Cray C90 (240 MHz)	16	20.65	13312	15
DEC AlphaServer 8400 5/612 (625 MHz)	40	20.54	24552	50
SGI Origin 2000 (195 MHz)	64	20.1	40000	25.0
Avalon	68	19.33	30464	72.5

Fig. 1. Some benchmarks for the comparison of currently used computers and supercomputers and a BEOWULF cluster of 70 nodes (AVALON) reported from Sterling et al. (1998).

4. Planning

The rationale for the planning follows the fact that the clusters of computers show actually the best price to performance ratio for a wide range of scientific and industrial applications. This is mainly due to the trend in the computer hardware market as it evolved in the last few years. Among the various cluster configurations proposed up to now for gross grain parallel computing the BEOWULF approach looks as one of the most promising. The BEOWULF approach is based on a distributed computing network realised by means of a server and a number of client nodes connected together through a fast LAN and usually operated in a Linux parallel software environment. Such clusters show a typical price as low as 20% of the price of an equivalent proprietary supercomputer, are fully modular, and allow one to track easily the fast evolution of the computer hardware. The CPUs can be replaced by new faster units without replacing the whole system at every update of the hardware. The BEOWULF technology is naturally fitted to fast processing large images such as those generated by the new mosaic CCD detectors. The current wide field CCD cameras used in astronomy use mosaics of CCD detectors of up to 4k×4k pixel. This structure implies that every image consists of blocks (extensions) obtained from physically independent detectors. This data structure, named embarrassingly parallel, is particularly suited to parallel processing since every extension can be reduced and calibrated individually before the recombination of the full sized scientific image. Great processing speed is required for such applications because several scientific programs allow for a maximum delay of 12 to 24 hours from the input of the raw data (100 GigaByte per day for OmegaCAM) to the output of the calibrated images. A benchmark done by ESO on a SUN Enterprise 450 workstation with four 250 MHz processors, 1 GB RAM, 9 GB hard disk and a 120 GB RAID disc and on a BEOWULF cluster with one server with Pentium III 550 MHz processor with 512 MB RAM and eight client nodes with Pentium III 450 MHz processor with 512 MB RAM and a 144 GB RAID disc resulted in a processing time of 35 minutes for the SUN and only 5 minutes for the BEOWULF cluster for the calculation of a masterbias image from a set of 12 WFI images of 8k×8k pixel each. The relative gain in the benchmark resulted greater than 7. This gain depends on the number and performances of the processors used in the cluster as well as on the performance of the

LAN supporting the cluster in terms of latency times.

The project has as one of its goals the study and optimisation of the connection between the server and the client nodes of the cluster. After the recent apparent superiority of the ATM technology for high speed networking, the most promising approach looks now to be the Gigabit Ethernet.

The implementation of a BEOWULF cluster using Gigabit networks is still to be studied in detail and it is likely that the use of such solution will require further changes in the communication protocols. Moreover, all nodes of a BEOWULF cluster must necessarily keep open channels (sockets) to input the requests from other nodes. This implies safety problems against illegal local or external access. The safety problem is particularly important if the BEOWULF cluster is managed not as a stand-alone isolated department facility but as one node of a geographical network distributed computing facility for national-wide access with dynamic allocation of the resources available. The bandwidth of the geographical network to be used for this application must be appropriately large. In Italy this should be possible through the use of the newly planned GARR-C network.

The realisation of the BEOWULF clusters planned within this program is first committed to processing mosaic images from astronomical wide field cameras already available, such as the ESO WFI camera of 8k×8k pixel format, or in course of realisation such as the OmegaCAM camera of 16k×16k pixel format.

The planning of phase I of the national program (months from 1 to 12) is based on the co-ordinated work of all four local research units and the rationale summarised above.

The implementation of the BEOWULF cluster (16 nodes) will be studied mainly by the research units at Capodimonte Astronomical Observatory and Trieste Astronomical Observatory (8 nodes) due to their greater availability of manpower and facilities suited to support the program.

The network related problems will be mainly studied by the research unit at Naples Federico II University with a BEOWULF cluster of 4 nodes.

The development of the application software for mosaic images will be done mainly by the research units at Capodimonte Astronomical Observatory and Trieste Astronomical Observatory.

The work will use the experience gained within the collaboration with ESO for the realisation of the photometric pipeline of the WFI camera. The Capodimonte Astronomical Observatory will also contribute to the program its availability of ESO VLT and VST guaranteed observing time obtained through its share into the VST OmegaCAM project. The ability to process WFI and OmegaCAM images will be used either for the future VST-300 Mpc Survey and for the definition and later scientific use of the observing programs on the large new technology spectrographs available on the ESO VLT telescope. This application will be developed mainly by the research unit at Trieste Astronomical Observatory due to the experience gained in this field through its collaboration with ESO on the UVES, FLAMES and GIRAFFE projects.

Further applications of the BEOWULF cluster will be done in the field of speckle astronomy due to the collaboration of the local research unit at Trieste University with researchers at ESO. Parallel BID processing of sub-images of size consistent to the isoplanatic range should allow to restore images wider than the isoplanatic range by means of a mosaic oriented approach.

The planning of phase II of the national program (months from 13 to 24) is similar to the planning of phase I and continues the activity scheduled hereby.

Based upon the full modularity of the BEOWULF systems, the research unit at Capodimonte Astronomical Observatory will purchase and install other 16 nodes plus 1 spare in order to reach a configuration with a number of nodes equal to the number of extensions of OmegaCAM (32). The performance gain of the full sized cluster will be evaluated for data reduction applications and the criteria for optimum distribution of the computational load on the nodes will be studied. Further study on the network tuning will be done based on the first benchmarks obtained on the prototype cluster realised in phase I of the program. The starting configuration of the network will be based on a 100Mb/s link between the client nodes and the switch and a 1 Gb/s link between the server and the switch.

Concerning the application software, the development of the BEOWULF pipeline software for the reduction of VST OmegaCAM astronomical images will be completed looking forward to the future VST-300 Mpc Survey. The study of the network connectivity and safety problems will be completed by the research unit at Naples Federico II University in co-ordination with the other research units using the prototype BEOWULF cluster with 4 nodes realised in phase I of the program.

The study and tests on the interaction of the CPU/network on the prototype BEOWULF cluster with 8 nodes realised in phase I of the program will be completed by the research unit at Trieste Astronomical Observatory. The astrometric pipeline software will be finalised. After the completion the BEOWULF system and the application software running on it will be used to process the preparatory fields required by the planned GTO observations at FLAMES and at other observing programs of the research unit.

The research unit at Trieste University will study the preliminary BID codes tested on scalar systems on the BEOWULF cluster with 8 nodes realised by the research unit at Trieste Astronomical Observatory in phase I of the the program. Tests on the 32 nodes BEOWULF cluster realised by the research unit at Capodimonte Astronomical Observatory in phase I and II of the program will also be done. Full test with simulated images and with true astronomical observations will be carried out before release of the application software to scientific use.

5. Organisation of the work

The organisation of the work in phase I of the program is the following:

(a) Research unit at Capodimonte Astronomical Observatory (Naples): purchase, installation, configuration and test of the prototype BEOWULF cluster with one server and 17 clients realised with AMD processors; issue of one contract for a collaborator to the program; study of the connectivity related problems and of the optimum balance between bandwidth, computing power and data storage within the cluster on an an appropriate number of nodes; development of a pipeline application software for the reduction of VST OmegaCAM astronomical images able to exploit the parallel structure of the cluster looking forward to the future VST-300 Mpc Survey.

(b) Research unit at Naples Federico II University: purchase, installation, configuration and test of the prototype BEOWULF cluster with 4 Pentium clients and Gigabit Ethernet; issue of one contract for a collaborator to the program; study of the network connectivity and safety problems.

(c) Research unit at Trieste Astronomical Observatory: purchase, installation, configuration and test of the prototype BEOWULF cluster with 8 AMD clients; issue of two contracts for collaborators to the program; study of the system optimization (CPU versus network) in collaboration with the Capodimonte research unit; development of the astrometric and photometric pipelines for the analysis

Fig. 2. The 8 nodes BEOWULF test cluster developed at Capodimonte Astronomical Observatory and available for the phase I of the project.

of WFI OmegaCAM images paying special attention to crowded stellar fields; tests on ESO EIS survey fields.

(d) Research unit at Trieste University: purchase, installation, configuration and test of a BEOWULF AMD cpu to

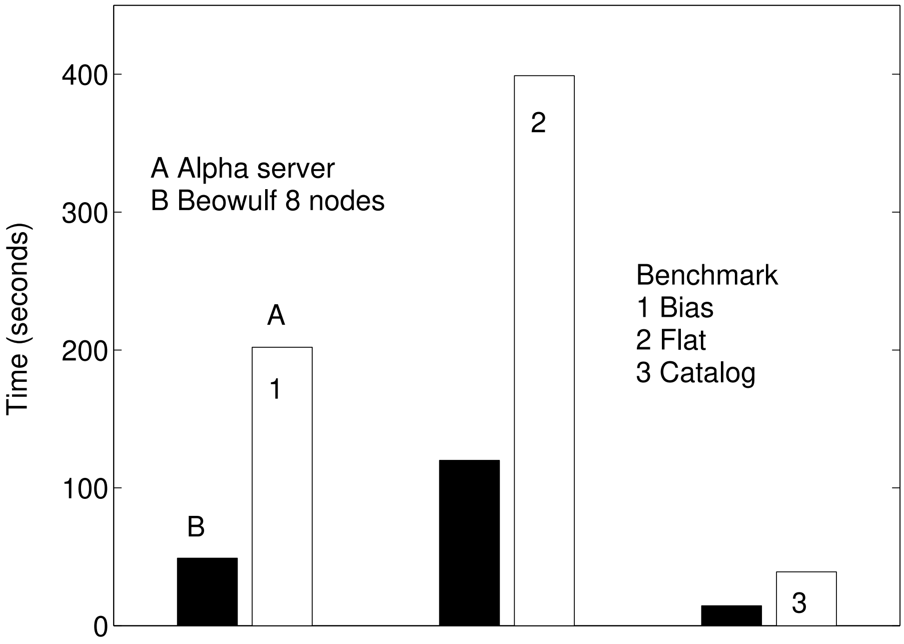

Fig. 3. Some timing results on typical operations on ESO WFI 8k images with the 8 nodes BEOWULF test cluster of Capodimonte Astronomical Observatory. The gain in computing sped of the BEOWULF cluster is self-evident.

be used as supervisor of the BEOWULF cluster at Trieste Observatory; issue of one contract for a collaborator to the program; preliminary parallelisation and integration of BID codes on the BEOWULF platform; preliminary tests on simulated high spatial resolution data.

6. Project evaluation criteria

The work done within the project will be monitored as per the COFIN evaluation criteria, documented, and finally distributed to the scientific community. The criteria which will be used for the evaluation of the project are those used in the definition and the feasibility study of the proposal, namely:

(a) relevance to current astronomical observing programs;
(b) relevance to planned astronomical observing programmes;
(c) need of a national facility committed to processing large format and mosaic images;
(d) expertise in the co-ordination of information technology programs of national level;
(e) expertise in the application of information technology to numerical processing of astronomical images and data;
(f) expertise in the realisation of computer systems for handling and processing of astronomical images and data;
(g) availability of internal manpower for management and control;
(h) availability of external manpower for realisation;
(i) availability of facilities for hosting the program;
(j) availability of funds for carrying out the program;

(k) availability of feedback from the scientific astronomical end users;
(l) return in mosaic image processing;
(m) return in numerical image restoration;
(n) return in applied supercomputing, network computing and handling of images data bases.

Further criteria suggested for the evaluation of the two phases of the program, on top of those used in the definition and in the feasibility study of the proposal, are the following:

(a) consistency and level of integration of the phases of the national program with local programs;
(b) co-ordination in time of the local phases with the national phases.

7. Expected results

The main results expected from the project are the following:

(a) Research units at Capodimonte Astronomical Observatory: Realisation, optimisation, and operation of a BEOWULF cluster with 32 client nodes and realisation and operation of the pipeline software for the reduction of VST OmegaCAM images for the future VST-300 Mpc Survey.

(b) Research unit at Naples Federico II University: Realisation, optimisation, and operation of BEOWULF clusters with 2 and 4 nodes with high speed LAN connection and completion of the study of the network connectivity and safety problems.

(c) Research unit at Trieste Astronomical Observatory: Realisation, optimisation, and operation of a BEOWULF cluster with 8 client nodes and realisation and operation of the astrometric pipeline software for processing the preparatory fields required by the observations programs of the research unit.

(d) Research unit at Trieste University: Migration and operation on simulations and real data of a mosaic version of the IDAC and bi-spectrum BID codes on the BEOWULF cluster with 8 nodes of Trieste Astronomical Observatory and test on the 32 nodes cluster realised at Capodimonte Astronomical Observatory.

Acknowledgements. The proposal of the BEOWULF project described here originated from a discussion with M. Capaccioli, Director of INAF Capodimonte Astronomical Observatory, Naples, Italy.

References

Boden, A. F., Redding, D. C., Hanisch, R. J., & Mo, J. 1996, JOptA, 1996, A 13-7, 1537
Jeffries, S. M., & Christou, J. C. 1993, ApJ 415, 862
Mancini, D., Sedmak, G., Brescia, M., Cortecchia, F., Fierro, D., Fiume Garelli, V., Marra, G., Perrotta, F., Rovedi, F., & Schipani, P. 2000, SPIE, 4004, 79
Ragazzoni, R., Marchetti, E.,& Valente, G. 2000, Nature, 403, 54
Reschke, C., Sterling, T., Ridge, D., Savarese, D., Becker, D., & Merkey, P. 1996, in High Performance and Distributed Computing, IEEE Computer Society, 626
Ridge, D., Becker, D., Merkey, P., Sterling, T., Becker, D., & Merkey, P. 1997, in IEEE Aerospace Conference
Roggemann, M. C., & Welsh, B. 1991, Imaging through Turbulence, (New York: CRC Press Inc)
Sterling, T., Becker, D. J., Savarese, D., Berry, M. R., & Reschke, C. 1996, in International Parallel Processing Symposium, IEEE Computer Society, 104
Sterling, T., Becker, D. J., Warren, M., Cwik, T., Salmon, J., & Nitzberg, B. 1998, in First NASA Workshop on Beowulfclass Clustered Computing, IEEE Aerospace

Valdes, F. 1998, in Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, et al. (San Francisco: ASP), ASP Conf. Ser. 145, 7

Valdes, F., & Tody, D. 1998, SPIE, 3355, 28

Warren, M. S., Becker, D. J., Goda, M. P., Salmon, J. K., & Sterling, T. 1997, in Parallel and Distributed Processing Techniques and Applications, 1372