A compilation and analysis of the data management and computing in the OPAL experiment, 1989 to 1996.

J. Baudot [a], C. Bondila [a], D. Charlton [b], D. Deatrich [c], L. del Pozo [d]. J. Letts [c], E. McKigney [e], S. O'Neale [b], M. Schröder [a], A. Williamson [f]

[a] *CERN, CH 1211 Geneva 23, Switzerland*

[b] *University of Birmingham, Birmingham B15 2TT, United Kingdom*

[c] *Indiana University, Bloomington, Indiana 47405, USA*

[d] *University of California, Riverside, California 92521, USA*

[e] *University College London, London WC1E 6BT, United Kingdom*

[f] *University of Alberta, Edmonton AB T6G 2N5, Canada*

We present a quantitative review of data collection, reconstruction, simulation and analysis in the OPAL collaboration. The evolution of the sample sizes, code and computing resources will be presented. The present system will be analysed with modern tools for modelling and metrics which are now available. [a]

[a] More details can be found at http://opalinfo.cern.ch/Opal/offline.html

Key words: OPAL; batch; interactive service; SHIFT; Monte Carlo; code management

1 The Data Flow

In OPAL, the data reconstruction and analysis computers are closely integrated with the data acquisition system [1]. Events are first recorded on disk at the filter [2], which is composed of a cluster of VME-based HP-UX workstations. At all stages, subsequent data transfers take place over local area networks (LAN) using disk buffers and permanent storage, which allows almost complete automation of the dataflow chain.

The compressed raw data is written to optical disks loaded into jukeboxes at the experiment. As well, it is written to tapes in the CERN computer centre over the LAN. A workstation farm at the experimental site reconstructs the raw data as it is collected, using the OPAL standard reconstruction program ROPE. This "Rope farm" consists of HP PA-RISC workstations. When the farm is not being used for reconstruction, it participates in Monte Carlo Production. There are currently 14 CPUs, with a capacity of 375 CERN units.

After a loose physics selection the DSTs are transferred over the LAN to the CERN computer centre, where they are recorded permanently on a large disk array in the SHIFT system[3]. These data are also backed up on magnetic tape media in the CERN computer centre. The data is accessible for the users about two hours after it has been collected. In general the data is not exported to outside institutes.

2 SHIFT - Scalable Heterogeneous Integrated Facility

Essentially all subsequent analyses of the data are performed on the SHIFT batch system, developped by an OPAL/CN collaboration. This system consists of two SGI Challenge XLs, providing 780 CERN

Preprint submitted to Elsevier Preprint 4 March 1997

Year	1990	1991	1992	1993	1994	1995	1996
Int. Lumi. / pb^{-1}	7	14	25	33	58	33 [6]	1 [20]
Events / 10^6	4.0	9.0	15.0	21.3	27.4	21.0 [2.7]	1.6 [8.4]
Hadronic events / 10^6	0.15	0.35	0.77	0.72	1.62	0.74	0.03
Raw Data size / GB	59	96	186	333	406	338 [50]	39 [200]
[D]DST size / GB	6.2	15.2	32.7	59.3	93.7	63	[88]
SHIFT Disk Space / GB	20	100	190	340	680	880	1100

Table 1
Number of events and sample sizes. Values in [] for running above the Z

Year	1991	1992	1993	1994	1995	1996
CERN Units in SnOPAL			92.8	243.7	418.7	737.7
Events from OPAL workstations	583748	1166491	1376634	3078492	7233124	11336976
Events from central facilities	199133	440383	1577264	2871995	2855856	3973298
Events from outside CERN	337642	550156	568452	1555783	6607794	5319079

Table 2
Monte Carlo production : SnOPAL capacity and number of events produced

Units of computing power and over 1 TB of disk storage. Files are mounted with NFS and read via RFIO.

Data from the highest reconstruction pass for each year and popular Monte Carlo samples are stored on disk. All other data are stored on tapes and catalogued by Fatmen. User analyses typically proceed by passing through the full DST data sample, producing ntuples or micro-DST summaries for further study. The permanent storage of the DST data on disk allows the analyses to proceed efficiently by directly accessing specific events using lists of file offsets (dadlists). Users can also make their own dadlists whenever appropriate, reducing the I/O and CPU needs for the analysis jobs.

SHIFT uses a modified version of NQS as batch system. Among the additional features added are dynamic load balancing and queue ordering. Jobs can be submitted from within CERN (80%) and from remote sites outside of CERN (20%), with log files returning to the original site, and other output files wherever specified. A project to test LSF as an NQS replacement is under way.

3 Interactive Services

OPAL uses a cluster of HP workstations, called SnOPAL ("Snakes on OPAL") for most of the interactive work and some batch jobs with little network I/O. Free CPU capacity is used for Monte Carlo Production.

The cluster has two boot node/file servers, which provide redundant copies of user applications. Because of the importance of stability for the Monte Carlo Production, the cluster is constructed in such a way that all nodes can continue running Monte Carlo jobs, even if both boot nodes are down. It is not possible to use the nodes interactively while the bootnodes are down, however, because some critical software is served.

There are currently 90 workstations in the cluster, of which 46 are used for Monte Carlo Production, the CERN Units shown in table 2 apply only to these.

IT Division is developing an "interactive benchmark" to model the typical interactive load on the CERN Work Group and Public Login Unix Servers. This model has been used to analyse the OPAL users profile on HPPLUS.

The readily available Unix accounting data is taken as the starting point for the model. The model groups commands into a small number of classes and an analysis of usage patterns for a 5 weeks period on HPPLUS shows three different categories of usage for OPAL. We interpret these as physics analyses, general usage and idle users. As a comparison, the ATLAS user analysis identifies a fourth category of usage, code development.

The differences in the usage between ATLAS and OPAL users are probably due to the fact that most OPAL users have accounts on SnOPAL, where they do the more CPU intensive tasks and code development.

Users/Server	Gen. Activ.	Idle	Analyses	Code Dev.
ATLAS/ATLAS_WGS	65%	12%	15%	8%
OPAL/HPPLUS	76%	16%	8%	-

Table 3
The proportion of users found in different categories

4 The Monte Carlo Production

The OPAL Monte-Carlo Production (MCP) system [4] is designed to manage the processing of generated events through the OPAL simulation package. It is organised around a central process which generates and distributes computing jobs to exploit free cycles on SnOPAL CPUs or the ROPE farm and dedicated simulation facilities. DDST generated offsite are shipped on DAT or DLT as international network bandwidth is inadequate for such transfers.

The control system was developed in Rexx. In 96 it has been moved to UNIX, and is now implemented in PERL. A WWW interface is used to get the physics needs and to provide documentation about the production status.

Whereas the system performs well, it has become quite complicated and inhomogeneous due to its development on different architectures with different languages. An attempt to re-design the system using object oriented modelling techniques (Booch-OMT) and a graphical tool (StP) has started.

5 The Code Management

OPAL created coding conventions early on for all reconstruction and analysis software. The purpose of these conventions is to ensure strict FORTRAN77 compliance, except for a mandatory 'IMPLICIT NONE' and MIL bit functions. Such conventions ensure that OPAL software works properly on all supported platforms.

The standard compliance of the code is verified by using compiler options and ftnchek on HPPLUS. Additional testing of the software is done on an SGI machine, with the help of compiler options for debugging. Recently we have started to use logiscope for measuring the quality of our code and identifying routines which might be difficult to test or maintain.

References

[1] J.T.M. Baines et al., Nuclear Instrum. and Meth. A325 (1993) 271-293

[2] D.G. Charlton, F. Meijers, T.J. Smith, P.S. Wells, Nuclear Instrum. and Meth. A325 (1993) 129-141

[3] B. Panzer, CHEP92, p. 483

[4] S. O'Neale, CHEP92, p. 471