ATLAS DB Meeting 17 March 1999

Harry Renshall

IT Division

Operational Report on Atlas 1998 1TB Milestone

Operational Objectives in 1998

To populate 1TB of Atlas simulated events into Objectivity data bases in a single federation and manage tertiary storage of these data into the High Performance Storage System using a SLAC/CERN (ASD group) interface through the linkable Objectivity Advanced Multithreaded Server.

Events were not to be generated but read from an existing pool of simulated events stored in Zebra format on magnetic tapes.

Performance was not an important part of this milestone (except as it affected the completion date) the emphasis being on functionality and operability.

Hardware/Software

- ATLOBJ01 - Sun Ultra 5, 300MHz monoprocessor, 64MB memory later expanded to 192MB, 2 disks of 23GB (1 disk crashed the system and could not be used) , fast ethernet. Ran Objectivity V4 linkable ams with hpss interface and an associated migration/purge server controlling a disk pool of 20GB on a single disk.
- ATLOBJ02 - Sun E450, 300MHz dual processor, 512MB memory, 2 disks of 18GB, fast ethernet. Ran Objectivity V4 linkable ams with hpss interface and an associated migration/purge server controlling a disk pool of 40GB on two disks.
- LOCKATL - Sun Ultra 5, 300 MHz monoprocessor, 64MB memory and mirrored external disk of 9 GB. Ran Objectivity V4 lock server and ams and hosted the federation lock and FDDB files.
- HPSS1D01 - IBM RS6000/F50, 266 MHz dual processor, 256MB memory, 120GB disk for HPSS user tapes and Atlas milestone, fast ethernet (up to 10 MB/sec full duplex). Connected to 4 STK Redwood tape drives on Dec Alpha servers via hippi (up to 100 MB/sec) and controlled a pool of 50 tapes of 50GB each.
- RSBATxx - Fifteen RSBAT nodes of 60 Cern Units and with fast ethernet and eighteen RSBAT nodes of 30 CU with fddi networking. Typically five were in use from this pool at a time running the Atlas application.

  The application read zebra events from a stage pool on an sp2 node over the CORE FDDI network, converted the events into persistent objects (ten times per input event) in databases that were either on ATLOBJ01 or 02 using the HPSS version of the AMS.

Performance Limitations

Hardware limitations include cpu speed on the RSBAT machines, disk speed (a single disk can read or write a single stream at 10 to 11 MB/sec - reading and writing at same time in the ATLOBJ01 configuration dropped to 2 to 4 MB/sec), and network connectivity. The total bandwidth out of the fast RSBAT machines (all applications) was 10 MB/sec. The total bandwidth out of the slow RSBAT machines plus many other machines on the CORE FDDI network, including the Zebra file stage server, to the ATLOBJY machines was 10 MB/sec.

These constraints meant very variable performance from individual batch jobs, worse performance (factor of two) from the slow than fast batch nodes and worse performance (factor of two) from the single-cpu, single disk AMS server than the double-cpu double disk server. The aggregate throughput to tertiary storage under the final stable running conditions was 1.5 MB/sec.

Operational Timeline

Effective data flow from batch jobs on RSBAT (LSF batch system) started around 11 Dec. to the ATLOBJ01 machine using a single disk instead of the two planned. Data rate observed (any number of jobs) was only 200 KB/sec and the machine was observed to be paging heavily mostly due to the parallel rfio processes with their associated memory buffers moving data to HPSS (typically 8 processes of 5MB).

On 14 December 128 MB memory was added and the throughput (several batch jobs) increased to about 500 KB/sec. A problem now was the single 23 GB disk which ran at about 1.5 MB/sec when only writing but dropped by a factor of three when writing and reading i.e. in the steady state.

On 18th December an optimised program version (one third less cpu usage) was ready and also the ATLOBJ02 server was started with data spread over two disks. This server delivered 1.5 MB/sec under optimum steady state conditions of three batch jobs using its AMS while ATLOBJ01 was delivering 0.5 MB/sec with two batch jobs. Probably the lower overall average of 1.5 MB/sec observed was due to some batch running on the slower batch machines. Unattended computer centre operation started on the evening of the 18th and the 1TB milestone was passed at 21.00 on 1 Jan 1999 (i.e. not in 1998 but before CERN officially restarted !).

The duty cycle during the unattended 10 days (CERN and Atlas staff checked progress daily from home or office) was only about 50%.

Stoppages in the Data Flow

- Two stoppages due to running out of batch work - one because the automatic afs token extension of the workflow manager process failed (not known why) and one because of (excessive ?) festivities. Total of 3 days lost.
- Four stoppages due to locks held by no longer existent batch jobs. This resulted in new batch jobs going into a permanent wait state. Two were due to 'normal' system component failures, two are unexplained (one of them was an Objectivity error 'expected page x, found page y' which I think is now fixed). Two lock situations were cleanly recovered by oocleanup -dead, the other two replied 'oocleanup already called' (for robustness the application called oocleanup at beginning and end) and had to be fixed by stopping/restarting the lock server thus killing the five running (waiting) batch jobs. This caused pain as they were run in dependent chains. Two days lost.
- One stoppage due to a series of tapes not being in the IBM robot. Fresh work from STK tapes was submitted while waiting for the tapes to be put in the robot.
- One crash of the afs server hosting the RD45 home directory. Although the executables were always copied locally the lock recovery tools became unavailable and we were in a locked situation. Three days lost.
- Two crashes of the HPSS data server - one hardware (transient), the other unknown. Since the HPSS data mover currently has to be restarted manually one crash resulted in the pool disks of both ams servers filling up and batch jobs dying leaving messy locks. Two days lost.

Conclusions

Most of the interruptions would have happened even with attended operations but they would have been fixed within hours not days. We must clearly improve the overall reliability of the distributed systems we have. In practise the AFS server hosting RD45 has not crashed since (others have) nor has the HPSS data server (though a second one has).

The HPSS software and internal performance behaved as expected with no interruptions, disk to tape running over hippi at tape streaming speed (8 to 11 MB/sec) and tape hardware compression of 1.6 on this data so only 14 tapes of 50GB were used.

The AMS/HPSS interface and migration purge/server proved robust.

The main cause for operational concern is the ease with which a federation can become locked and the complexity of recovery. The blunderbus approach we used could easily have corrupted parts of the federation.

As regards performance we can only say the setup proved itself to be highly non-optimised both in throughput and recovery during the unattended operations.

Tests with Compass show an AMS on a machine like ATLOBJ02 but which is properly networked and has multiple disk strings can sustain 7 MB/sec throughput (gigabit ethernet input to the AMS and gigabit output with rfcp to tertiary storage).

We have started migration of the computer centre servers to a clean switched fast ethernet/gigabit ethernet backbone, due to complete by end March. We have already replaced the disks on ATLOBJ02 with a modern higher data rate RAID-5 array and hope to repeat this milestone (or at least a significant fraction of it) quite soon.