Computing and Analysis Model

E. Buckley-Geer

- Present hardware model
- Present data analysis model
- Parameters for Run II system
- Assumptions for Run II
- Possible Approaches

Present hardware model

The present model aims to keep the data as close to the CPU as possible.
Note that even though we have significant computing in the trailers we do NOT provide high bandwidth access from those machines to the data that is on the central systems.
We have NO dedicated batch systems, all systems are used for both interactive work and batch processing.

* FNALD consists of older VAX CI nodes, 15 satellite Vax 4000/90's, and two Alpha 2100 servers each with two processors. The satellite and Alpha nodes are clustered with CI nodes via FDDI. FNALD FDDI ring is on one of four CDF FDDI Gigaswitch ports.

	CPU in VUPS	Data Disk	8mm (8500)
CI nodes	125	150 GBytes	12
FNALKx	405	230	22
Alpha	700	216	8

* CDFSGA system consists of an SGi Challenge XL with 24 200 MHz processors. CDFSGA is networked to one of four CDF FDDI Gigaswitch ports.

	CPU in VUPS	Data Disk	8mm (8500)
CDFSGA	2000	468 GBytes	16

(NOTE: Most of the data disk on CDFSGA is fast/wide SCSI and single stream jobs that are not CPU bound can read as fast as 9 MBytes/sec on striped disk sets. This is about a factor of 4 higher than rates achieved on other CDF systems.)

* B0/CDF Portekamps Computing

CDF portekamps contain several computing clusters consisting of Vax/VMS, Alpha/VMS, SGi, and IBM workstations. Clusters are all Ethernet based with bridges to FDDI backbone (one of four CDF FDDI Gigaswitch ports).

	# of Stations	CPU VUPS	Disk GBytes	8mm drives
local clusters	150	4600	600	120
VAX/ALPHA	114	3000		
SGI	36	1600		

(about 1/3 of machines are 25 VUPS or faster)
(about 1/5 of machines are 50 VUPS or faster)

Utilization for November

System	Total VUPS	VUPS Used
FNALD	900	1300 (69%)
CDFSGA	1600	1300 (81%)
UNIX CLUSTER	1600	200 (12.5%)
VAX CLUSTER	3000	750 (25%)

All systems are running a mix of interactive use, Monte Carlo generation and data analysis

Accounting for CDFSGA for December

Jan 5 11:48 1996 User report for December 1995 Page 1

TOTAL	1071360	---	kennedy	10979	1.02	rolli	2754	0.26
<unused>	486955	45.45	rebcdf	10552	0.98	worm	2594	0.24
paulini	130245	12.16	benlloch	9597	0.90	ryuichi	2397	0.22
paoti	101503	9.47	root	8837	0.82	pinghu	2263	0.21
ohmoto	62334	5.82	owen	8810	0.82	engels	2253	0.21
jeffc	47069	4.39	kuwabara	7403	0.69	asakawa	2042	0.19
rharris	31268	2.92	jpdone	5432	0.51	wasiq	2018	0.19
valls	27453	2.56	ashmansk	4225	0.39	tlthomas	1701	0.16
hskim	22847	2.13	tmiao	3967	0.37	takanot	1625	0.15
kordas	20205	1.89	xinwu	3451	0.32	tannenba	1506	0.14
niimi	17324	1.62	troconiz	3258	0.30	murat	1416	0.13
mdpeters	11167	1.04	brink	3089	0.29			

Jan 5 11:48 1996 Command report for December 1995 Page 1

TOTAL	1071360	---	build_jo	37604	3.51	wxmod2.e	8505	0.79
<unused>	486957	45.45	jetrad.e	31219	2.91	exclsv1b	8420	0.79
grandmc.	112317	10.48	lep_d_kp	20337	1.90	cdfsim.e	7472	0.70
pt_ana	101328	9.46	dbltag.e	13510	1.26	psigen.e	6378	0.60
inclsv1b	50040	4.67	blcharm.	10884	1.02	exclsv1a	5980	0.56
blife.dx	47011	4.39	dipho92	10490	0.98	wmod3.ex	5714	0.53

* About 25% of CPU used in December was used for Monte Carlo

Data Storage

Disk space allocated to Physics groups:

Physics Group	FNALD	CDFSGA
Monte Carlo	21	0
EWK	50	0
TOP	76	30
QCD	26	24
EXOTIC	33	32
BOTTOM	114	76

115 GBytes of Miscellaneous data space on FNALD (small groups, usr$scratch, LOOK areas, etc.)

20 GBytes of Miscellaneous disk space on CDFSGA

STK SILO
157 GBytes on CDFSGA of staging space
65 GBytes on FNALD of staging space

Many disks are NFS cross mounted between CDFSGA and FNALD. These mounts are not always reliable and we always try to KEEP THE DATA CLOSE TO THE CPU. Some data may be replicated on more than one system.

The STK Silo contains 6000 cartridges each with a size of 200 MBytes. We are upgrading to new drives and 800 MByte cartridges. The Silo has 4 drives on CDF-SGA and 4 drives on FNALD. Access to data stored in Silo is ONLY through FATMEN catalog and automatic staging system.

Most accessed datasets are stored on disk if small enough. Somewhat less frequent (or larger) datasets are stored in Silo. Largest or least accessed datasets are stored on 8mm tape only.

Note that we also make copies of some datasets for distribution to outside institutions. This includes some “library tapes” for use on site in the trailers.

8mm TAPE USAGE

	RAW	PROD	USER
usage rate	200	600	100 per month approximate use

Four categories of tapes.

RAW data tapes are written in B0 and transferred to FCC.

PRODUCTION data tapes are mainly written on the farms.

USER tapes are tapes available for physicists doing analysis.

COPY tapes - copied about 14000 tapes, about 3600 are for LIBRARY and TRAILER, rest go off site.

Present Data Analysis Model

- Raw data in YBOS banks. Data on 8mm tape.
- Reconstruction. Add physics information. Compress information to create PAD banks. Still in YBOS format. Split data into production datasets. All on 8mm tape.
- Physics Dataset creation. Performed by physics groups, e.g., creation of BPADS (includes compression of banks - 30% space saving), inclusive high-Pt lepton samples etc. Many man-months of work, lengthy to recreate. Can be small enough to fit on disk, depends on dataset. Still in YBOS format. Up to this point all data is file-based.
- Physics data selection. Create ntuple, keep small amounts of information, e.g., for high-sumet analysis, 250 bytes per event. Small enough to fit on workstation disk. Physics groups also make standard ntuples, e.g., photon, jet. Data now in ZEBRA RZ format.
- Mosts analyses start with PAD data.
- Small number require full DST information, e.g. W mass measurement.
- Ntuples heavily used
- For historical reasons PAD data and ntuples in different formats.

Parameters for Run II system

Reminder of what Rob presented at a previous meeting

- **Present L3 cross-section (stream B only) = 316 nb at 2$\times 10^{31}$ peak luminosity gives 6.3 Hz**
- **Scale by luminosity increase $\times 10$ and allow 60% for cross-section or increase in compute time $\Rightarrow$ factor of 16 increase over Run Ib.**

Run Ib data volume

- **45 M events $\times 130$ Kbyte $\Rightarrow$ 6 Tbytes**
- **DST 190 Kbytes $\Rightarrow$ 8.5 Tbytes**
- **Splitting 8.5 $\times$1.3 overlap factor $\Rightarrow$ 10 Tbytes**
- **PAD 32 Kbytes $\Rightarrow$ 1.4 Tbytes**
- **Split PAD $\Rightarrow$ 1.9 Tbytes**
- **Total $\Rightarrow$ 28 Tbytes**

Run II data volume - estimated from Run Ib

Assume 250 Kbyte event size, 600 Million events per year (about 1 fb^{-1}) (factor 16 increase)

- **RAW 136 Tbytes/year**
- **DST 200 Tbytes/year**
- **Split DST 260 Tbytes/year**
- **PAD 17 Tbytes/year**
- **Split PAD 23 Tbytes/year**
- **Total 640 Tbytes/year - present thinking in the CD has robots capable of about 100 Tbytes of storage.**

If we store each event once then we are talking of more like 200 Tbytes/year which is still more than would fit into the above size robot!

Assumptions for Run II

- Data will get from B0 to FCC without being carried there by a human. We will write directly to the mass storage from the DAQ staging disks.
- It is a good idea to keep the cpu that accesses the data as close as possible to the data store, i.e. if you submit a query to the data store, the actual query runs close to the data and just returns the results, which are hopefully a lot smaller than the complete dataset!
- Access to pieces of the event should be possible. Event storage is not file based any longer.
- Large robotic tape storage will be necessary. Will need to be expandable. Example, IBM 3494 holds 2800 tapes at 10 Gbytes/tape $\Rightarrow$ 28 Tbytes. Fall 1995 prices are about \$25/Gbyte (0.7M\$ for 28 Tbytes).
- Speed of tape media must increase. Example, take the low P_T electron dataset which will be 2 Tbytes/year in Run II. At present 8mm tape speeds (450 Kbytes/sec) it would take 50 days to read. We would like to read a primary dataset such as that in $\approx$ 1 day $\Rightarrow$ 20 Mbyte/sec tape speeds. If we had tape speeds of 100 Mbytes/sec we could spin through it in 6 hours. Note that this assume the cpu can handle that speed of data.

- **Large amount of disk for frequently accessed data. Stream A PADS will be about 2 Tbytes/year - should fit on disk. Presently we have 28 Tbytes of data and about 600 Gbytes of disk in the central systems. This is a factor of 47. Keeping this ratio we would need 10 Tbytes of disk for 2 fb^{-1}. Assume \$300/Gbyte $\Rightarrow$ 3M\$! So we need to shrink the event size and only store necessary information for the events on disk.**

Possible Approaches

CAP (Computing for Analysis Project)

- Research project in High Performance and Parallel Computing group in CD
- Assumes that users want to apply complex selections to data using a large set of event attributes, and that all the physics information (or even the entire dataset) is needed in order to select the events. The small set of chosen events is then transferred to a user workstation for further analysis. Selection is made efficient by organzing data into physics objects such as electrons, jets etc (object-oriented uses C++, objects are addressed using persistent pointers). Most selections involve only a few types of objects so the entire data need not be scanned.
- This approach is good for the final event data, e.g. PAD type data but is not applicable to storage of the raw data. So there will by necessity be different formats to store the data.
- We can arrange for someone from the CAP effort to talk to us if people are interested.

Integrated Approach (I-A)

- Integrates the data management early in the analysis cycle. Information about an event, whether from the RAW data, the DST etc is collected in a database system. Data organization spans levels of workgroup servers and local analysis platforms and each maintains its own database, retrieving missing info from a primary repository. Assumes that selection cuts are simple and change incrementally but that the user ultimately needs the full data for a few selected events. Uses knowledge of how selection criteria evolve to optimize the data organization.
- This is an idea of Stephan Lammel. He will talk about this at a future meeting.

Both these are described in a CHEP95 talk located at http://www-chep95.fnal.gov/abstracts/abs_96.html