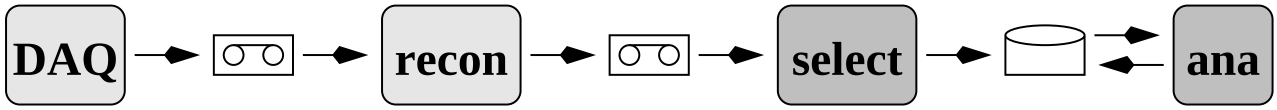
Run II Data Handling and Analysis Systems

Walsh Brown, Liz Buckley-Geer, Terry Watts, Eric Wicklund, and Stephan Lammel

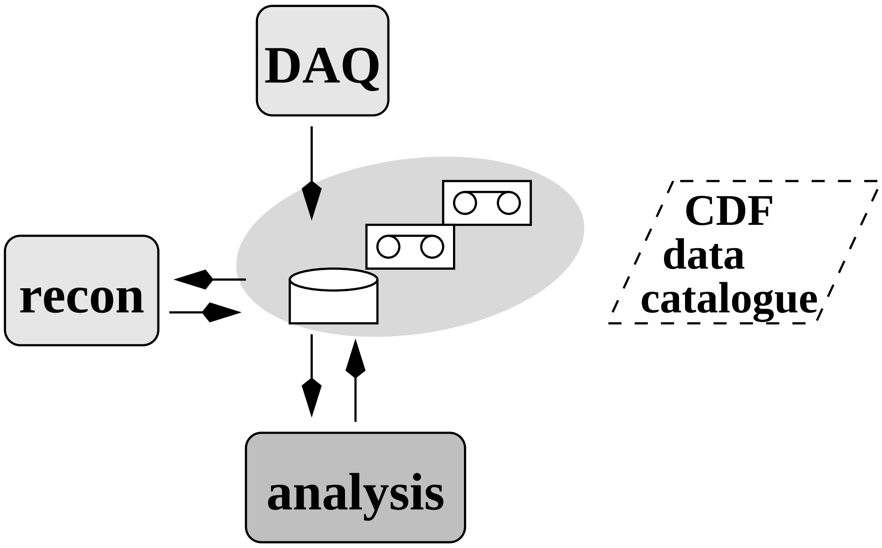
- **New Thinking ?**
- **DH Issues**
- **Design Goals**
- **CDF Data Catalogue**
- **Data Handling System**
- **Analysis System**
- **Concluding Remarks**

New Thinking ?

Data access should be simplified over Run I (i.e. no intrinsic knowledge of the data handling system should be required, what is where and how to find the right database information, etc.).



Data Access by dataset name. This means we can drop the data-flow thinking of Run I but think about data access from a pool of hirarchically managed storage.



Data Handing Issues

The main issues all relate to the information density. With the old-fashioned ROOT file based approach we have an intrinsically low information density:

- I/O bandwidth: we are interested in a tape technology where the incremental cost for additional bandwidth is affordable;
- disk space: efficient management of disk space is desirable;
- central system(s): offload work from the central systems as much as possible:
  - workstations
  - Monte Carlo productions
  - remote analysis/compute resources
- overall tuning: with the intention to maximize the disk space available to physics analysis;

Metadata information has to be managed externally to ROOT. A good integration with the CDF data catalogue is then necessary.

Design Goals

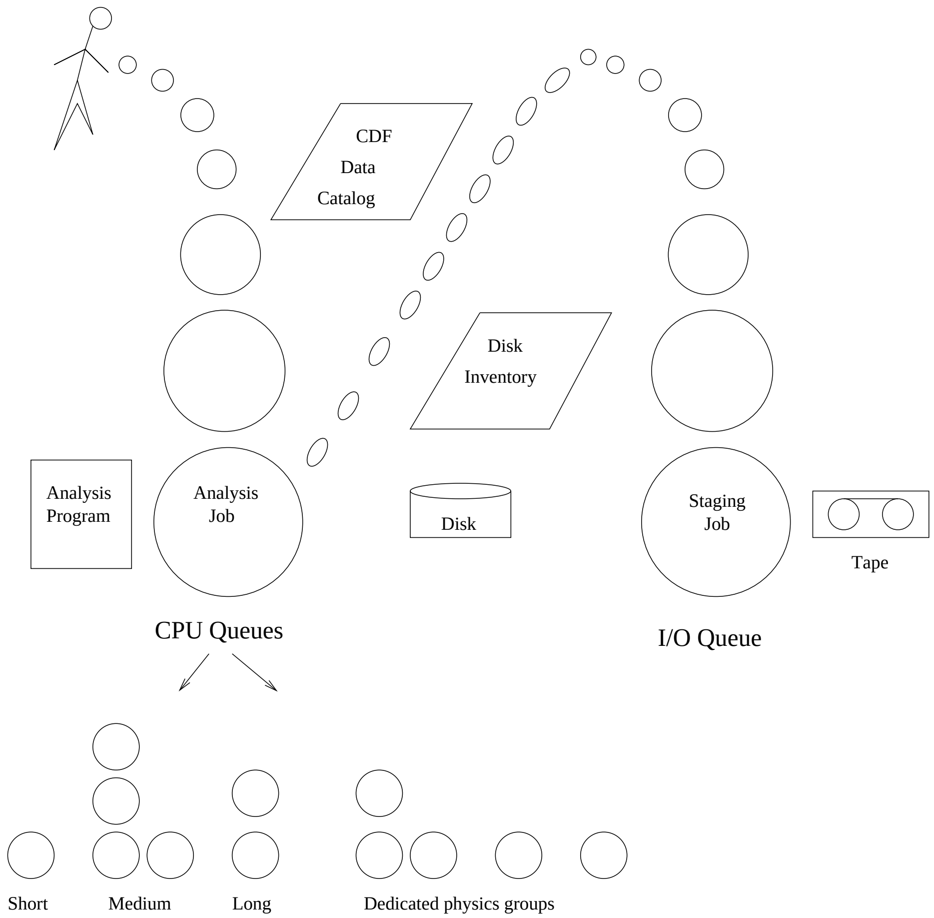
- Support as many analyses as possible;
- Support the physicist in her/his way of working and thinking
  (a system designed for the users and not users for a system);
- Scalable beyond current estimates
  (ability to integrate new technologies);
- Require as little operational support as possible
  (Stephan and Terry will be busy discovering SUSY and Higgs);
- Flexible, to shorten turn-around cycle during commissioning phase;
- Easy export to and installation at remote sites.

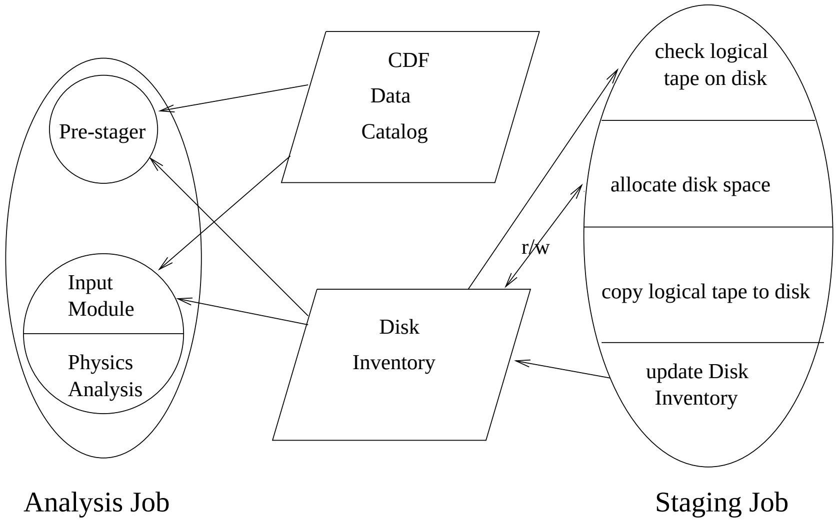
CDF Data Catalogue

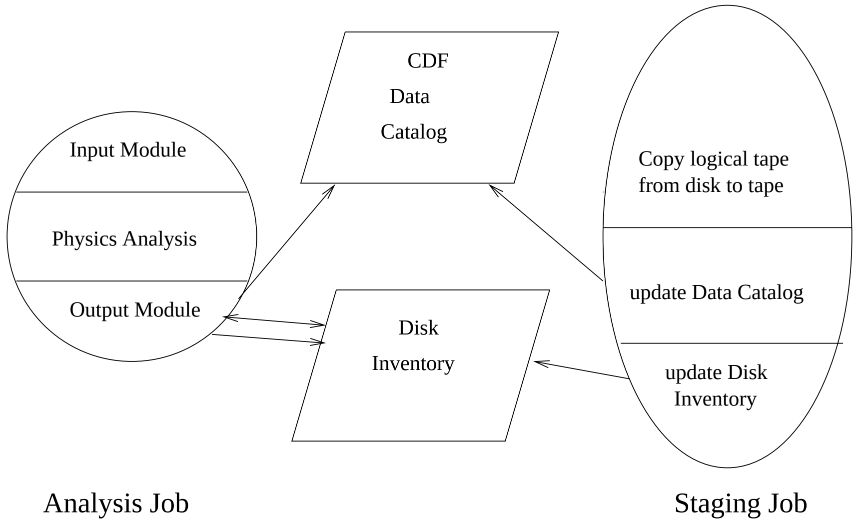
Eric is working on the CDF data catalogue. The database will hold all the metadata of the data that is now located in various different databases (production manager database, tapes database, luminosity database, etc.).

A professional DBMS will be used. Information will be organized in a hierarchy from datasets to file sets, files, and run sections.

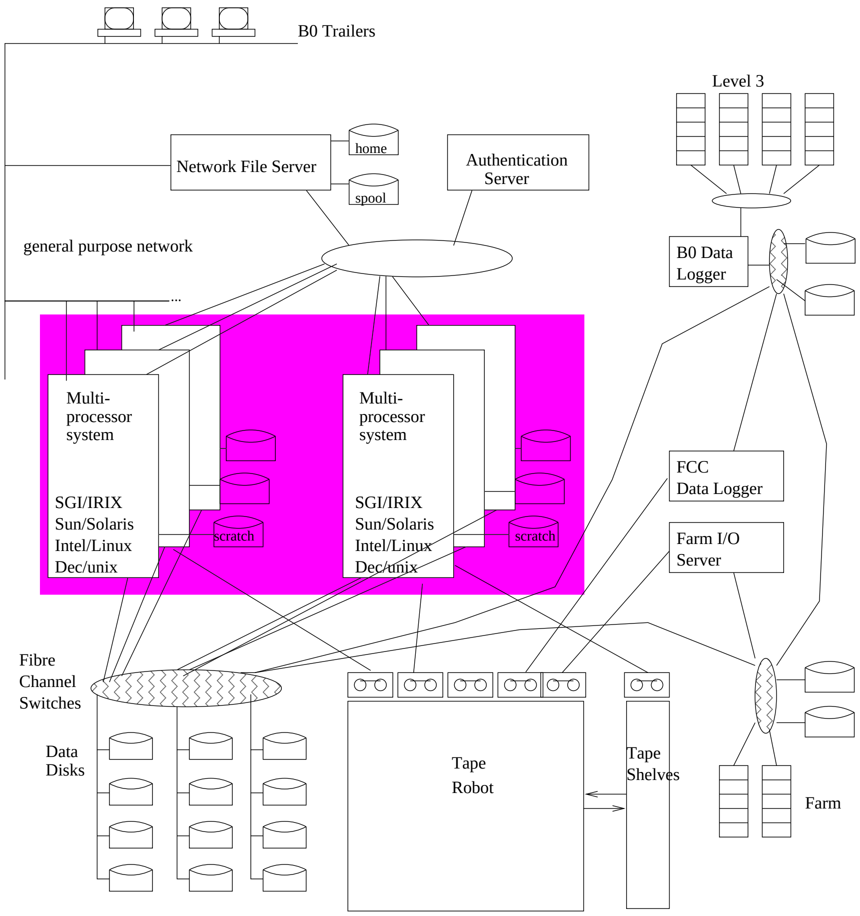
Data Handling System

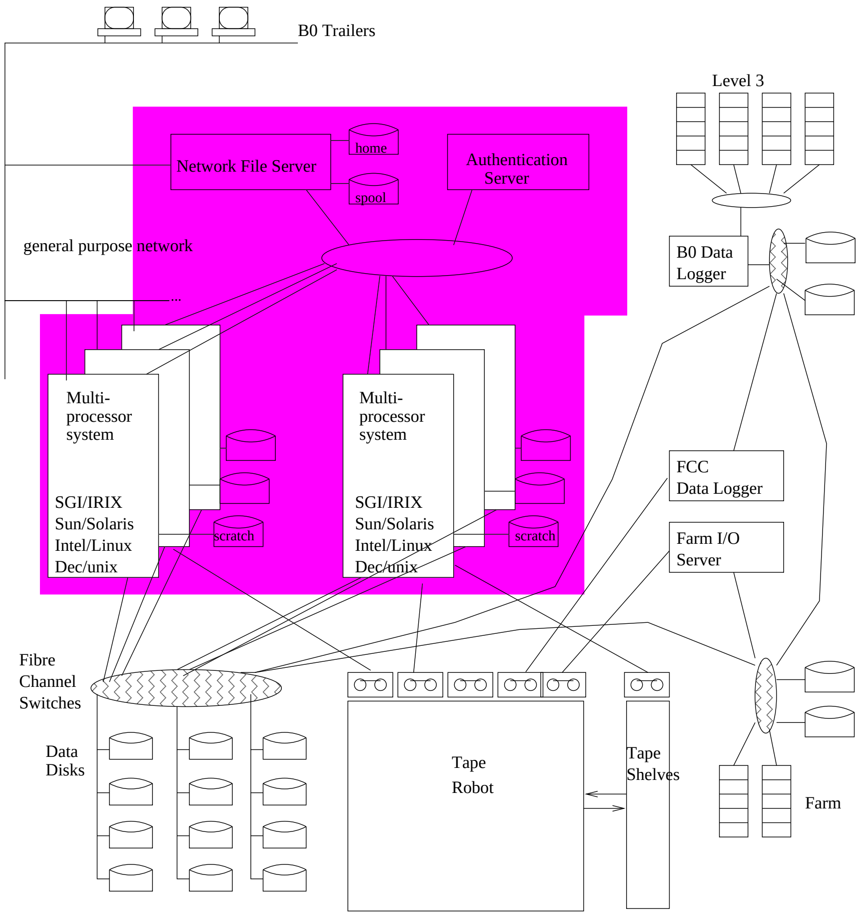


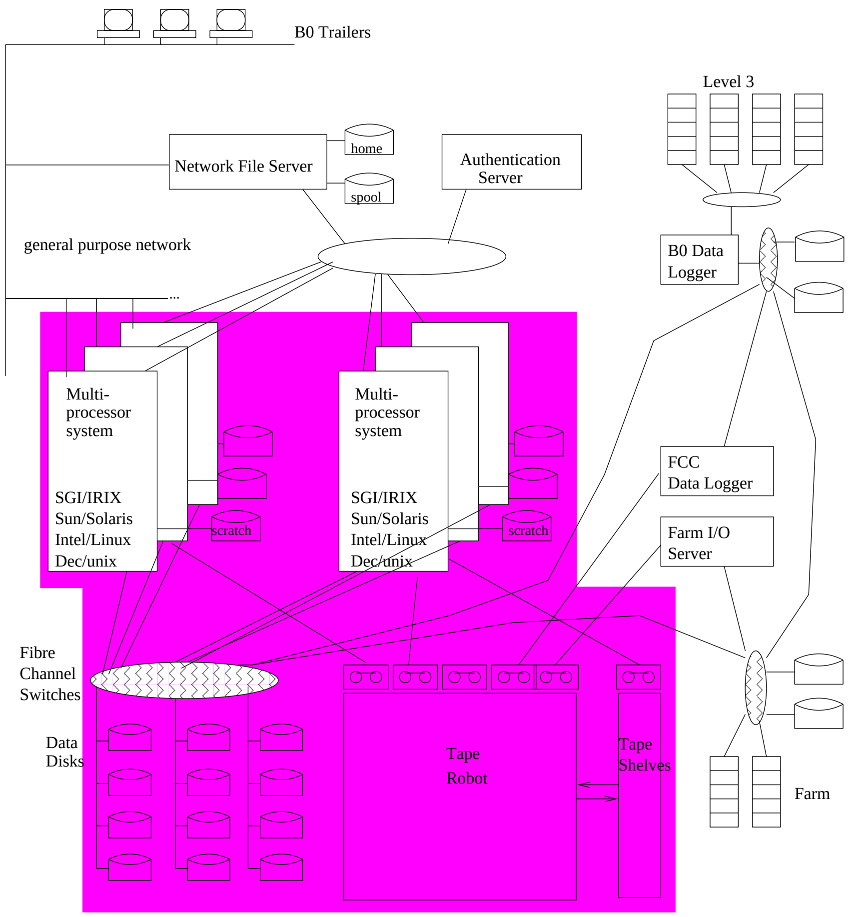

CDF
Data
Catalog
check logical
tape on disk
Pre-stager
allocate disk space
r/w
Input
Module
copy logical tape to disk
Disk
Inventory
Physics
Analysis
update Disk
Inventory
Analysis Job
Staging Job

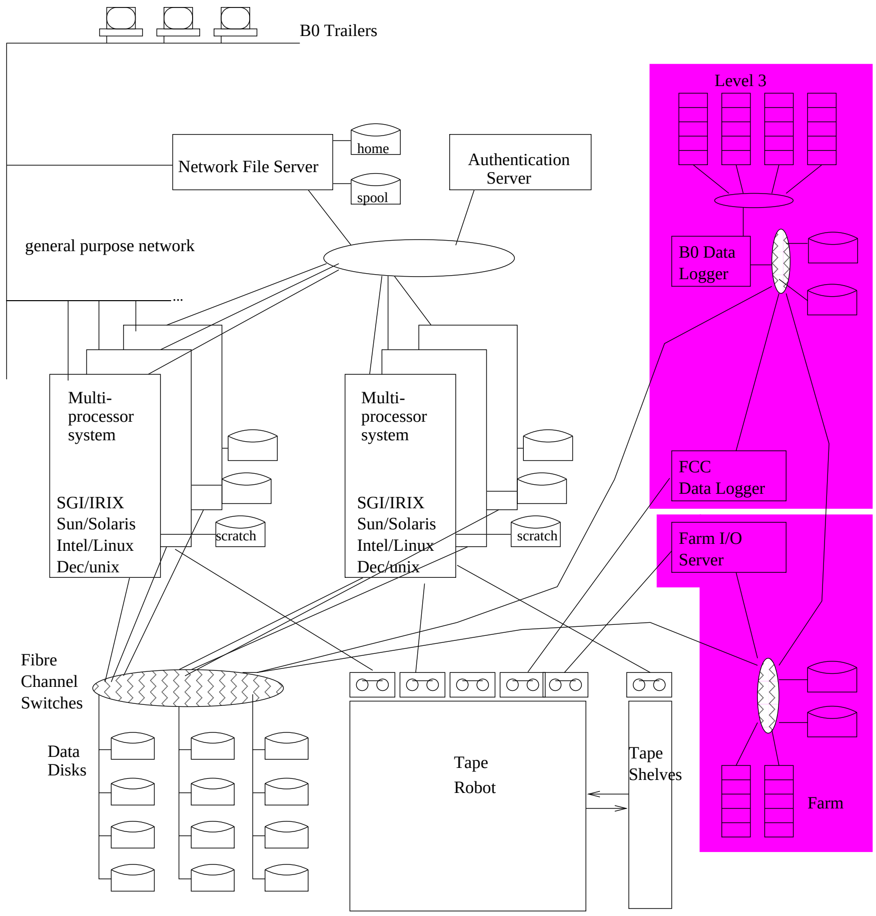

CDF
Data
Catalog
Input Module
Copy logical tape
from disk to tape
Physics Analysis
Output Module
update Data Catalog
Disk
Inventory
update Disk
Inventory
Analysis Job
Staging Job

Analysis System




B0 Trailers
Level 3
Network File Server
home
spool
Authentication Server
general purpose network
B0 Data Logger
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
scratch
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
scratch
FCC Data Logger
Farm I/O Server
Fibre Channel Switches
Data Disks
Tape Robot
Tape Shelves
Farm


B0 Trailers
Level 3
Network File Server
home
spool
Authentication Server
general purpose network
...
B0 Data Logger
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
scratch
scratch
FCC Data Logger
Farm I/O Server
Fibre Channel Switches
Data Disks
Tape Robot
Tape Shelves
Farm


B0 Trailers
Level 3
Network File Server
home
spool
Authentication Server
general purpose network
...
B0 Data Logger
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
Multi-processor system
SGI/IRIX
Sun/Solaris
Intel/Linux
Dec/unix
scratch
scratch
FCC Data Logger
Farm I/O Server
Fibre Channel Switches
Data Disks
Tape Robot
Tape Shelves
Farm

Status

There are two critical software

- We have started to build a small scale prototype of both the disk manager and batch system to convince ourselves (and others) with a proof of principle.
- The final version of the disk manager must run in a heterogenous environment and handle network attached disk (i.e. keep track of mount status, perform mount operations, etc.).
- We have a null version that can handle all the inventory and usage tracking.
- We have decided to use/try LSF as batch system for the first prototype. Liz and Walsh have installed and configured the first queues on our prototype platform.
- Overhaul and upgrade of tape reading/writing software is almost complete.

and one critical hardware component.

- A small network file server arrived a week ago for evaluation. We are in the process of setting it up.
- Fibre Channel test equipment was ordered in May but we are still waiting for missing equipment before we can do any testing/learning.

Schedule

- end of September: results from first DH prototype
- September/October: offline godparent review
- middle of October: CDF data handling at PAC meeting
- beginning of November: setup of tape library
- end of December: small scale analysis system
- January 1999: von Rueden review
- January 1999: start of online and farm integration
- March 1999: start of analysis system hardware procurement
- April 1999: selection of tape technology
- summer 1999: fully functional Run II analysis system

Concluding Remarks

- System for ROOT based file I/O worked out (through several levels of details).

- First pass on software and hardware architecture complete.

- Critical components/aspects are identified.

- Prototype of a small scale system is in progress.

- Schedule for prototype/WvR review is tight but Run II schedule has large contingencies.

- It's time to do some work (or ask questions if you don't know yet how to help)...