



# TESTING THE CDF DISTRIBUTED COMPUTING FRAMEWORK

V. Bartsch\*, M. Burgon-Lyon\*\*, T. Huffman\*, R. St. Denis\*\*, S. Stojek\*\*\*

\*Oxford University, Oxford OX1 2JD, UK

\*\*Glasgow University, Glasgow G12 8QQ, Scotland, UK

\*\*\*FNAL, Batavia, IL 60510, USA

## Introduction:

### CAF:

- wrapper around a batch system (FBSNG or CONDOR) to submit jobs in a uniform way
- submission to CDF clusters inside and outside Fermilab from many computers with kerberos authentication possible

### SAM:

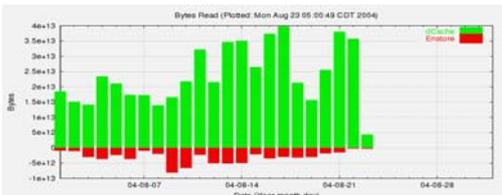
- data handling system of CDF in run II
  - ⇒ storages, manages, delivers and tracks processing of data
- designed to copy data to remote sites
  - ⇒ designed with remote analysis in mind

### Technical details:

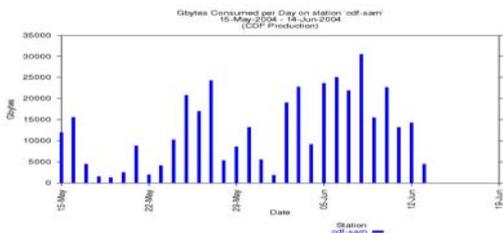
- major source of CPU power for CDF (Collider Detector at Fermilab) is the CAF (Central Analysis Farm)
  - ⇒ CAF: 800 nodes total CPU about 1200GHz
  - ⇒ additional: CondorCaf 400 nodes total CPU about 2000GHz
- outside Fermilab DCAF with a total CPU of about 1000 GHz and a disk space of 35 TBytes
  - access datasets which comprise a large amount of files and analyze the data.
  - autumn 2004 some of the important datasets will only be readable with the help of the data handling system SAM (Sequential Access to data via Metadata)
  - CAF and SAM have not yet been used in combination
    - ⇒ tests of the systems necessary

## Stress tests on the standard CAF at Fermilab:

Total amount of data daily read by CDF on the CAF



consumed data per day of the central SAM station during the tests



### Stress tests:

- create the usual user load
  - ⇒ 50 SAM projects on the CAF and move 20 TBytes per day
- split jobs in several segment
  - ⇒ run several parts of the job in parallel on different CPUs

### Limitations:

- submitting more than 100 SAM project at one time
  - ⇒ problems with the project master
- large number of files (order of 10 000 files)
  - ⇒ optimizer problems: checks location of files when requesting next file of the dataset or after network outage, blocks other requests
  - ⇒ could slow down all SAM station
  - ⇒ problem is already solved, but number of files request still should be small

### User friendliness:

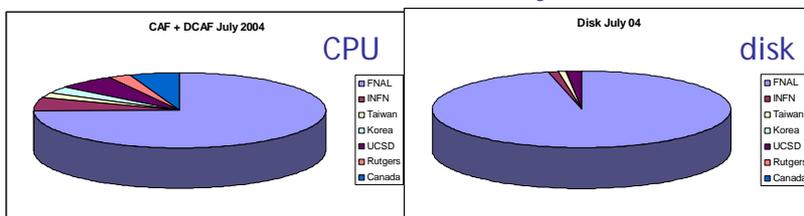
- recovery of partly failed user projects possible with `sam recovery dataset` command
  - ⇒ depends on the correct (or not at all) release of the file

### Conclusion:

- test phase successful, deployment of SAM for the CAF system foreseen this autumn

## Decentralized CAF:

Comparison of the total amount of CPU and disk space of the CAF and the DCAF systems



### Goal: 2005 50% of the computing outside Fermilab

- ⇒ distributed computing
- ⇒ use of DCAF (Decentralized CDF Analysis Farm)
- ⇒ SAM station environment has to be common to all stations and adaptations to the environment have to be