CDF Run II Computing Workshop.
A user's perspective

Stefano Belforte
INFN - Trieste

- http://www.ts.infn.it/~belforte + click on "offline page" + …
- top link on my web page on cdfsga
- CDF WinCenter: u:/belfo/slides/computing/ccw_10feb99.ppt

"A user of what ?" (Avi Yagil)

- Perspective of a User(s) of data at a remote institution
  - Data Analysis in Italy for CDF Run II

- Why ?
- I have to make a plan for computing for the Italian CDF collaborators for Run II:
  - what hardware
  - where
  - when
  - how much money

Which Hardware for data Analysis in Italy ?

- CPU, Disks, Tapes (Robots?), Network (LAN, WAN)
- I am going to share my exercise now with you
  - hopefully I learn something from discussion
  - maybe I help you focus your questions
- Italy is many institutions, all sizes, one will be like yours
- what is really different ? The WAN maybe.
  - But in RunII many transoceanic institutions….

Run I vs. Run II

Beware these numbers. Very difficult to make good predictions. Hope that conclusions do not change if numbers are a bit wrong.

	1994	now	2000	factor	note
DATA	100 inv. pb	-	2 inv. fb	x20	
TAPE size	2 GBytes	30 GBytes	50 GBytes	x25	$$
DISK size	2 GBytes	34 GBytes	100 GBytes	x50	cc
CPU	VAX3100	P2-500MHz	?	x100	ccc
LAN	10 Mbit/s	100 Mbit/s	100 Mbit/s	x10 (at best)	$$
WAN	1 Mbit/s ?	2 Mbit/s ?	4 Mbit/s ?	? x4 ?	S.O.S.
DISK I/O	5MBytes/s	14 MByte/s	20 MByte/s	x4 ?	S.O.S.

- Bottleneck: I/O, very difficult to put data into the CPU
- solution: bring the CPU to the data, build powerful cluster

Hardware at home. Copy FNAL setup again ?

- FERMILAB
- 500 GBytes → ≥20 TBytes
  ☞ x 40 !
- VMS cluster → high performance Fiber Channel based Storage Area Network
- Hand tapes + Silo → million $ robot

- PADOVA / PISA
- 30 GBytes → 2 TByte ?
- VMS cluster → Just a Bunch of Unix/linux boxes
- hand tapes → hand tapes ?

Anyhow, simple scaling doesn’t work. Data can not be parted among physicists.

Also would like to do better then Run I, more analysis more easily

How much data must I handle at home?

- Liz’s Table, see :
  http://www-cdf.fnal.gov/internal/upgrades/computing/offline_minutes/buckley_vr_talk_jan_99.ps
  Page 5.

PAD vs. Ntuple

- Ntuple: 1000 variables/event = 4KByte/ev (PAD = 60 KB)
- High Pt: ⇒ O(1% total data) keep all events
  - PAD: 2TBytes, 20 tapes ⇒ analyze at home
  - Ntuple: 200 GBytes ⇒ keep on PC hard disk
    - ☞ but ! Need several versions of the Ntuple, reduce 1/4 at most
- Low Pt: ⇒ O(10% total data)
  - PAD: 20TB, 200 tapes ⇒ have to do something
  - Ntuple: 2TB ⇒ don't fit on disk !
    - ☞ Reduce data sample ? 1/10th ⇒ Low-Pt = High-Pt
- anyhow is analysis dependent, ,many people... many opinions
- still… how do I bring those data in Italy ? Few tens to few TB..

Network needs for Analysis
(from a talk to INFN in Bologna January 19, 1999)

- 3 Scenarios (extremes and middle way):
  - copy all PADs in Italy, need almost no net to US
  - leave all PADs & "ntuple" in US, use Xterm from Italy
  - copy some PADs here, keep most Ntuples here (copy/create)
- Difficult to estimate requirements. Better the other way around. Given 4 Mbit/sec dedicated to CDF, what can we do ?
  - 4 Mb/sec = 200GByte/week = 2 tapes/week, can't beat DHL !
    - ☞ 1 tape a day = 100GByte/day = 10 Mbit/sec
    - ☞ **PADs don't travel on the net**
  - 4Mb/sec / 10 users = 1GByte/5hr/person for copying Ntuples
    - ☞ one analysis = 1/10th data ⇒ PAD=20TB, Ntuple=20GB ? refreshing Ntuple takes 4 days min ! More data, more users..
    - ☞ Converging argument: 10GB ntuple/physicist = minimum !
    - ☞ **Can't make Ntuple offsite and copy locally on the net**

What goes out of FNAL ?

- PADs don't travel on the net
- Ntuples don't travel on the net
- what do I do ?

What is an Ntuple anyway?

- Do we really need to refresh ~200 GBytes of Ntuples "continuously" ?
- The Ntuple is what we use for interactive histogramming.
  - If it takes one hour to get the histogram, may very well submit a job and get the histograms back.
    - ☞ Data transfer is limited, it makes no difference where the job runs !
  - An Ntuple is a data-set you go through in a few minutes at most
- disk $\rightarrow$ CPU: 50 Mbytes/sec at most $\Rightarrow$ 3 GBytes/min at most.
- Ntuple will always fit in your desk(lap)top PC !
    - ☞ Notice: Run I equivalent (200MByte) required good chunk of 5" big SCSI disk !

Disk to CPU

- beware the 3 GBytes/minute
- Getting data into the CPU is the bottleneck of today's computing
  - ➢ standard desktop PC vs. Power Unix Server
  - ➢ CPU may not be the limit anymore
  - ➢ Easy to have more CPU at home than at FNAL, very hard to have better I/O
  - ➢ could take 10 times longer to run "PAW" at home than at FNAL... (PIAF…?)

Getting Ntuple home

- **Easier way:** the WAN. T1 for 2 (500Kbit/sec/user) $\Rightarrow$ 0.5GByte/hour (now we deal with CDF notes, a few MBytes, here we go x1000 !)
  - ➢ 6 hours to get my Ntuple, a day or two more likely… NO WAY !
    - ☞ internet future may be brighter, Ntuples may be bigger…
    - ☞ <u>if possible:</u> maybe slow, likely unsafe, but easy
- **3 alternative solutions:**
  - ➢ Don't do it!
    Run "PAW" at Fnal (just Xterm+telnet). Fast, easy and safe
    - ☞ 500Kbit/s = 10 good Xterm sessions (or 5 perfect)
  - ➢ FedEx (1 lb., 5 days a week) easy and safe
    - ☞ Fnal -> US 1st(2nd) day: 400(200) $/year
    - ☞ Fnal -> Europe/Japan/Taiwan: 7 K$/year
  - ➢ create Ntuple locally from FedEx'ed PADs safe but **hard**

Data Analysis in Pisa: The Final Answer
(from a talk to INFN in Pisa May 12, 1998)

- We will have to try, can't pick the right approach before collaboration has finalized data handling and distribution tools, and analysis topics have been pinpointed
- We will try everything, user pressure will drive
- Needs will be dominated by physics output (maybe we find SUSY in 3-lepton samples and everybody looks at this small data set…)
- We will exploit local computing as much as possible to reduce network load (likely bottleneck, as it always has been)
- Still will need to access FNAL PADs to produce data sets to copy to Pisa. If network is no good will use tapes (expensive though!). But we **desperately need guaranteed bandwidth for interactive work**
- **If can not log in FNAL, no way to do most analysis here, only use "dead" data sets: no express data, no hot topics, just late sidelines… the good old way: take the plane and go to FNAL.**

Dealing with PADs at home

- Easily 1~5 TB, what to do ?
  - All on disk ? 20 to 100 disks… and no VMS cluster...
  - Multi CPU Server+RAID ?
    - ☞ Small SUN ? Or Big PC ?
  - PC farm (our small Level 3?) ? LAN !
  - Tape stacker ? 1~2 TB only? A couple of drive ? So slow !
  - Taking shift at tape mounting ? 5 PCs, 10 drives, 50 tapes… but… will beat the robot ?
  - Power server at FCC ?
- Up to 500GB: all on disk. 2 ~ 3 PC's working together.
  - ☞ LAN, LAN, LAN !
- Morale: the less you need it, the better it is.

Where to put a Power Unix Server with 5 TB disk ?

- See e.g.:

  http://www-cdf.fnal.gov/upgrades/cdfdh/
  doc/hardware/hard_arch.ps

  a.k.a. cdfsga:/cdf/pub/cdf4707_r2dh_hard_arch.ps

  Figure 2

Tapes

- Anyhow will need tapes for more then import to disk
  - PADs, simulation, MonteCarlo, …
- Will need to run analysis jobs from tape, just like at FNAL
- But in Run II all tape data must be spooled to disk first
- spool space: 100 GBytes each (one full tape) ?
  - Not likely
  - Better spool 10 GBytes a time
  - Better be a free parameter

Summary

- need FedEx
- need to run at FNAL: low latency WAN
- need flexible software at home
- need good LAN at home
- need flexible expandable hardware at home

Conclusions

- Recommendation to INFN

- Recommendations (requests) to managers

To my colleagues

- buy “nothing” now
- buy little (few PC’s, few tapes, little disk) next year, add disks as needed (JIT !)
- get the best LAN and WAN you can
- try to do the most at FNAL, ship tapes every day if need be, put CPU and or disks in FCC if needed
- see how it goes, see where the market goes
- be prepared to handle few TB in 2001/2
  - ☞ get a computer room ready
  - ☞ don’t know which hardware will be best, but likely it will not fit on your desktops

To the Offline Managers

- tapes, disks & CPU for everybody (lot’s of)
- friendly, low latency batch i/f from home (www ?)
- fast, easy-to-use i/f from Robot to FedEx
- help for simple linux system at home:
  - ➢ suggested/supported hardware configuration
  - ➢ easy to use/install software, adaptable to limited hardware setup
  - ➢ one example on-site OFF the LAN