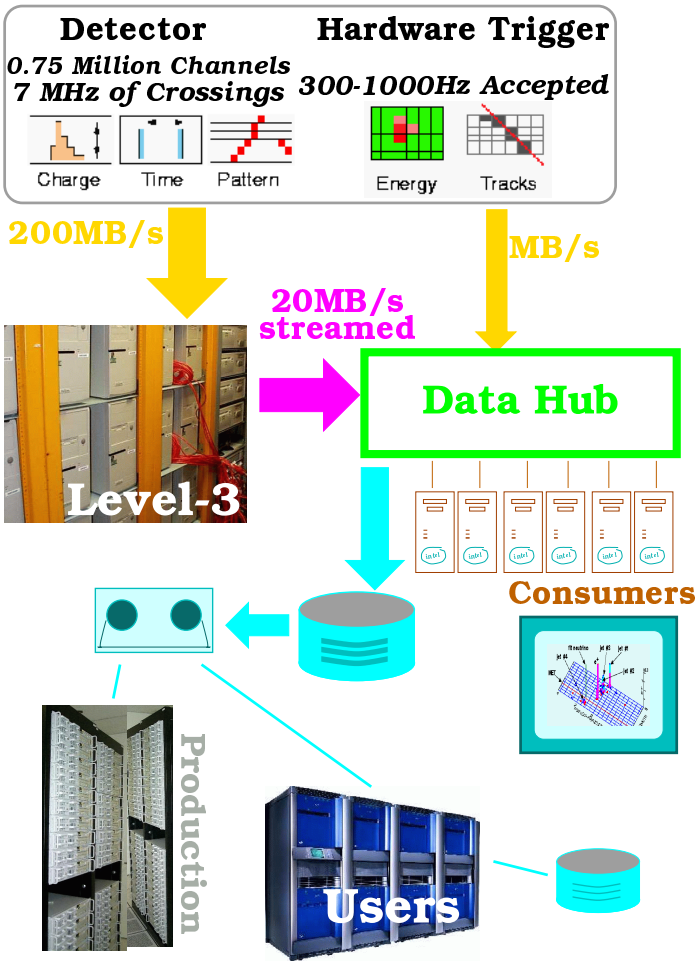
Data Handling Requirements

Robert Harris
Kevin McFarland
CDF Offline Operations
November 19, 2001

---

1. Ground Rules
2. Subsystem Needs
   - Raw data archiving
   - Production Farms
   - Data skimming
   - User analysis
3. Sum of Parts

Ground Rules: The Subsystems



- Level-3 Output (physics runs) is all archived
  - ↪ Data Hub (CSL) separates into $\mathcal{O}(10)$ streams
- Production output is split more finely, all archived
- Skimming may keep small fraction of data
- User Analysis comes in many varieties
  - ↪ Re-running CPU intensive analysis (e.g., tracking)
  - ↪ Or simple running through N-tuples

More Ground Rules

- Run IIa only ($2\mathrm{fb}^{-1}$)

Year	Peak $\mathcal{L}$ (10^{32}/cm^2/s)	Int. $\mathcal{L}$ (/fb)	
2001	0.82	0.5	
2002	1.34	1.6	
2003	1.51	2.3	$\longleftarrow \approx$ Run II spec
2004	2.27	4.1	
2005	3.26	6.9	
2006	3.91	10.5	
2007	4.95	15.0	

  - ↪ Peak rate for IIb is thought to reach $2.5\times$ IIa (luminosity scaling)
  - ↪ Integrated rate for IIb is $7.5\times$ IIa (luminosity scaling)
- Specified Rates relative to Peak
  - ↪ Always archive at peak rate
    - ⋆ This assumption is built into sizing the CSL buffer
  - ↪ Farms must be able to keep up with weekly peak
  - ↪ Other systems can operate at rate set by yearly average

Yet More Ground Rules

- Formats
  - ↪ Raw data is 250 kB
  - ↪ PADs (after commissioning) are 100 kB
  - ↪ Skimming may remove significant numbers of events
    - ⋆ E.g., inclusive high p_T electrons are ultimately ~ 2 Million events, but will log ~ 50 Million candidates
  - ↪ N-tuples may be very small
    - ⋆ "Standard" N-tuples perhaps 30 kB
    - ⋆ "Specialized" (e.g., TOPFILL) N-Tuples ~ 10 kB
- Some analysis aspects will be regimented
  - ↪ Skimming is now being coordinated by physics groups
  - ↪ N-tuple creation may be common as well
  - ↪ Unfortunately (or fortunately, depending on point of view), even priority physics analyses lack "WBS-level" of specification
- Much analysis will be done off central systems
  - ↪ Network export is very common now
  - ↪ Desktop disk and CPU is very cheap

Raw Data Archiving

- CDF data logging: < 90 Hz peak of 250 kB raw events
  - ↪ Over a year, average/peak $\sim \frac{1}{3}$
  - ↪ Over a week, average/peak $\lesssim \frac{2}{3}$
  - ↪ CSL buffer sizing can always write to tape at peak rate
  - ↪ Our record weekly peak is currently 3 TB/week, about 30% of specification
- Require “express stream” (3%) to LOOK area

Production

- Ultimate production output is 60–100 kB PADs
  - ↪ Now, output is 200 kB gzip'd RAW+reconstructed data
- Stated requirements:
  - ↪ Handle DC rate of 2–$3 \times$ "DC logging rate"
  - ↪ Keep up within $\mathcal{O}(1)$ day
  - ⇒ Require 30 MB/sec input
  - ⇒ Require 12 MB/sec output
- Is this matched to CPU now?
  - ↪ Current farm size is ≈ 500 PIII/500MHz CPUs
  - ↪ Analysis time is $\mathcal{O}(10)$ sec (same unit)
  - ⇒ 12 MB/sec in required
- Current operations. . .
  - ↪ Are output limited, only achieving peak of 5 MB/sec (2.5 Million, 170 kB events per day)
    - ⋆ Previously worse because of limited output buffer disk
      (we have increased to 2 TB since last review)
    - ⋆ Now partly because of limited tape drives
      (have 8; would add 4 more if could handle operational load)
    - ⋆ Now largely because of operational problems in tape writing

Skimming, Secondary Datasets

- This is where it all gets *fuzzy*. . .
  - ↪ These estimates rely on Savard and Bellforte's presentations at October CAF workshop
  - ↪ Should this committee recommend WBS level analysis specs from physics groups? Would it serve a purpose?
- Dataset sizes for $0.2\mathrm{fb}^{-1}$
  - ↪ Top/EWK requests 25 Million events in 2ndary datasets
  - ↪ QCD "satisfied" with these datasets – Savard
  - ↪ *Assume* exotics adds a similar load (excluding data sets shared with Top/EWK)
  - ↪ Bottom. . . well, assume need every event from Level-3 hadronic B, lepton+displaced track, Ψ
    - ⇒ 100 nb or 20 Million events
  - ⇒ Total output of 70 Million events
    - ⋆ Assume current event size with all RAW data, 170 kB (compressed)
    - ⇒ Requires 12 TB
  - ↪ Assume datasets are made twice from each event out of production
    - ⋆ It's much less now
    - ⋆ Remember that we are assuming that we re-run production with $\sim \frac{2}{3}$ of the farm CPU
    - ⇒ Lifetime of production output is only $\mathcal{O}(6)$ months when we have $2\mathrm{fb}^{-1}$

Skimming, Secondary Datasets (cont'd)

- Assume no caching on input to skim jobs...
  - ↪ But that caching satisfies 90% of subsequent requests to read these datasets!
    (Can cache 75% of this data with DIM-managed disk on 'sgi2 today)
  - ↪ As we go to $2\mathrm{fb}^{-1}$, assume this drops dramatically ($\sim 15\%$)
    - ⋆ But could be mitigated by more disk

User/N-tuple Analysis

- 2ndary datasets will be frequently exported, $\mathcal{O}(5)$ times
- N-tuplizing as a separate step
  - ↪ For small 2ndary datasets, analysis will proceed on PADs (almost certainly cached or exported)
  - ↪ For larger 2ndary datasets, analysis will proceed from N-tuples
    - ⋆ Assume each N-tuples made many times, $\mathcal{O}(5)$
- Assume N-tuples are either cached on disk or are loaded from tape to be exported ($\mathcal{O}(10)$) times
- Bottom line: there are lots of guesses here!
  - ↪ Nothing *a priori* unreasonable however

Summary of Requirements

Run IIa Requirements ($2\mathrm{fb}^{-1}$)

	Input (MB/s)	Output (MB/s)
Raw Archiving	n/a	22.5
Production Farm	30	12
Skimming/Striping	24	4.5
User/N-tuple	50	2.5
Total	105	43

- Disagree with the assumptions? Change them! `http://www.pas.rochester.edu/~ksmcf/dh`
- Largest item, cache misses in N-tupling, would be reduced if dramatically expanded ($\rightarrow 60$ TB) DIM disk
- Provides a basis for sizing final Run IIa system
- DAMNAG approximately predicted DH I/O needs

Summary 2002, $200\mathrm{pb}^{-1}$ Scenario

Run IIa Requirements ($200\mathrm{pb}^{-1}$)

	Input (MB/s)	Output (MB/s)
Raw Archiving	n/a	22.5
Production Farm	30	24
Skimming/Striping	41	1.5
User/N-tuple	3	0.5
Total	75	48

- What is different?
  - ↪ PAD is assumed (pessimistic) to still have all RAW data
  - ↪ Existing DIM disk only assumed
  - ↪ "More urgent" need to analyze data (less time)
- Note that bottom line doesn't change much
  - ↪ Impact of large PAD looms in distance!
- Summary of needs in this model:
  - ↪ 13 TB of DIM and Physics group N-tuple space should cache *all* latest 2ndary datasets & N-tuples
  - ↪ Bandwidth needed is dominated by farms
  - ↪ Export bandwidth needed (DC) is 11 MB/sec

Summary Run IIb

Run IIb Requirements ($15\mathrm{fb}^{-1}$)

	Input (MB/s)	Output (MB/s)
Raw Archiving	n/a	56
Production Farm	75	30
Skimming/Striping	60	17
User/N-tuple	190	10
Total	325	113

- What is different?
  - ↪ DIM and N-tuple disk pools are $\mathcal{O}(100)$ TB
  - ↪ "Less urgent", datasets can take longer to analyze
    (lifetime of 2ndary dataset is one year)
- *Roughly*, it scales with instantaneous luminosity under these assumptions