D0 Run II Data Management and Access

CHEP 97

April 4-11 1997

Overview

- Run II Data Management Needs
- Run I Experience
- Run II Design Considerations

DATA WARS

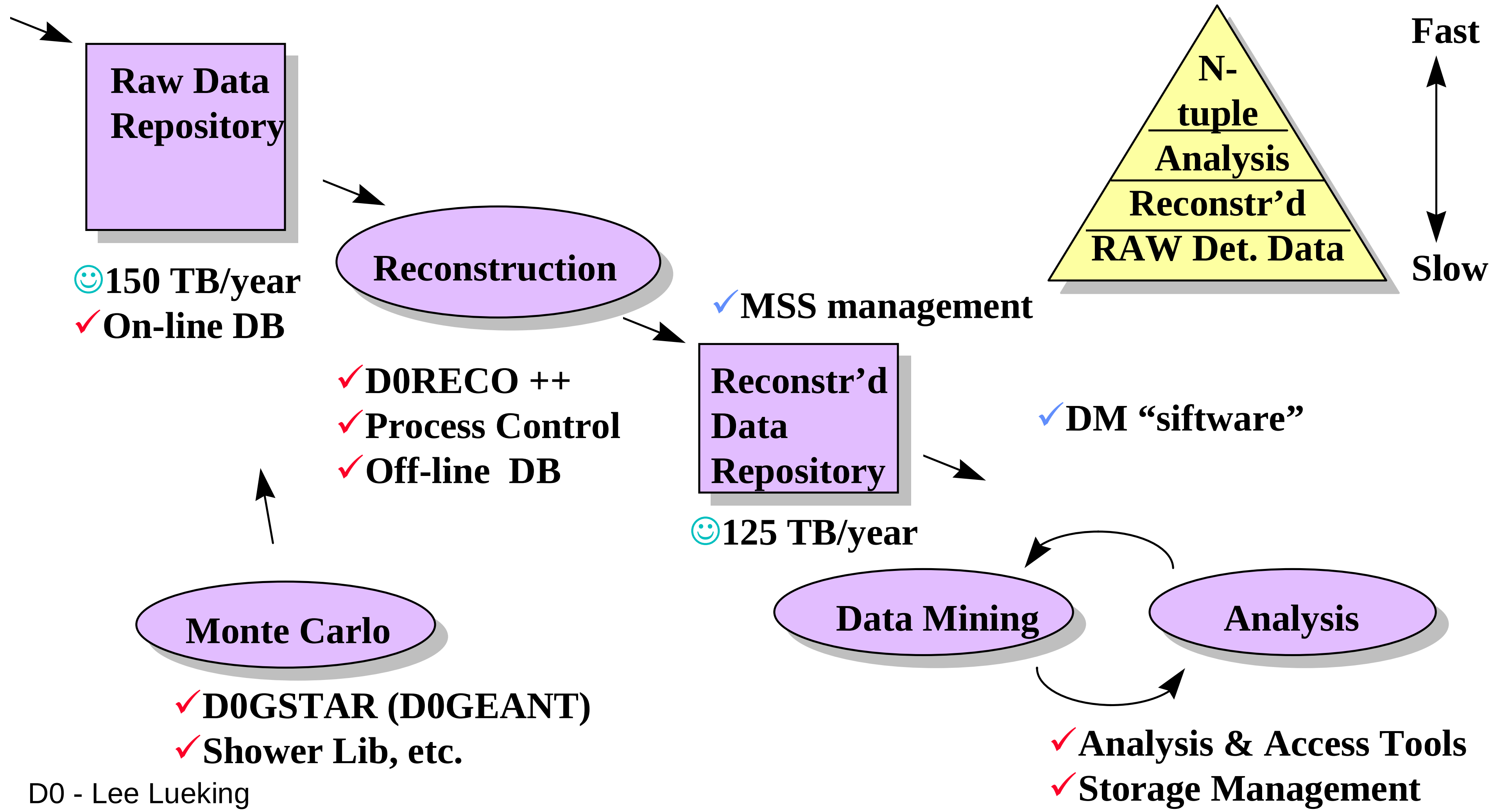
Episode II. Evil forces from the DAQ side threaten the on-going storage of knowledge. The data Management Empire strikes back with a series of attacks employing technical solutions and streamlined approaches to data processing and storage.

Summary of Run II Physics

- **Top Physics**
  - ttbar Production Cross Section
  - Top Quark Mass
  - Top Quark Decay Properties
  - W Polarization from Top Decays
  - Single Top Production
  - Search for the Charged Higgs Bosons in Top Decays
  - Discovery of New Physics in the ttbar Channel
- **Electroweak Physics**
  - W Boson Mass
  - W Boson Decay Properties
  - W Boson Charge Asymmetry
  - Forward-Backward Charge Asymmetry of Z Decays
  - Anomalous Vector Boson Couplings
- **QCD**
  - Parton Distributions
  - Next-to-Leading Order and Beyond
  - Drell-Yan (W/Z) Production
  - Jet Algorithms and Studies of Jet Structure
  - Rapidity Gaps, Diffractive Scattering and the Pomeron
- **B Physics**
  - Bs Mixing
  - CP Violation
  - Other B Physics Topics
- **Searches for New Phenomena**
  - SUSY -gaugino,stop,squark-gluino…
  - W',Z'
  - Many others...

Nearly Every area is statistics limited!

Off-line Overview



D0 Run II DAQ Rates

Parameter	Value
Peak Rate	53 Hz
Average Event Rate	20 Hz
Event Size	250 kB
Average Data Rate	5 MB/s
Level 2 Output	800 Hz
Level 3 Rejection	40

- Goal is to achieve 2 fb^{-1} recorded Luminosity over 2 year period.
- Factor of 40 in Level 3 is still being studied (optimistic).

D0 Run II Data Storage

Parameter	Value
Events	600M
RAW	150 TB/year
Reconstructed	75 TB/year
Analysis	50 TB/year
Thumbnail	3 TB/year
Total	280 TB/year

- Assuming significantly reduced storage per event compared with Run I.

D0 Run II CPU

Parameter	Need
Reconstruction/event	1000-2500 MIPS
Total reconstruction	30,000-80,000 MIPS
Analysis	60,000-80,000 MIPS

- Reconstruction CPU is based on:
  - » Run I reconstruction time per event and,
  - » C++ version of fiber tracking code being tested for Run II.
  - » Run I experience for farm efficiencies.
- Analysis CPU is scaled from Run I by a factor of 20 (the increased number of anticipated Run II events).

Run I Experience

- Data Set Sizes
- Access Patterns
- Major Constraints

Run I Dataset Sizes

Data Type	Size/event	Total Size	Comment
RAW	500 kB	30 TB	Raw data
STA (reconstd)	250 kB	15 TB	Re-reco-able
Streamed STA	250 kB	10 TB	10 x5% streams
DST(reconstd)	50 kB	2 TB	semi-re-reco-able
Streamed DST	50 kB	3 TB	14 streams
mDST	5 kB	0.2 TB	Virtual streams
ntuple etc.		0.3 TB	Analysis spec.
Total		55TB*	42 M events

- ***This sample represented about half of the total Run I storage.**

Run I Data Access

- Provided access to data by:
  - » Tiered data types: RAW > STA > DST > μDST.
  - » Physically (STA & DST) and virtually (μDST) Streamed data sets.
  - » Random access staging facility to get STA and DST events.
- Search (eg Top, SUSY)
  - » Sift through DST or μDST to find most interesting events.
  - » Extract or "pick" these events, based on run # and event #, from STA .
- QCD, b-Physics
  - » Go through DST data 2 to 3 times per year to make ntuples
  - » Use ntuples for all analyses.
- ElectroWeak (W Mass)
  - » Employed STA data stream. Studied, reconstructed, et cetera.

Run I Major Constraints

- Time for Access
  - » Typical time to process all μDST was ~2 weeks.
  - » 1 to 2 months to access information was considered a major, but manageable task. Typical for DST samples.
  - » Tape access was, in general, slow and tedious.
- Disk Space
  - » Data samples were designed to match available disk space.
  - » The complete μDST sample was 200GB, many local subsets of STA,DST and μDST data were also on disk.
  - » Never enough!
- Software Configuration
  - » Careful preparation and testing always gets in the way of progress.
  - » Streamlining obvious bottlenecks would help.

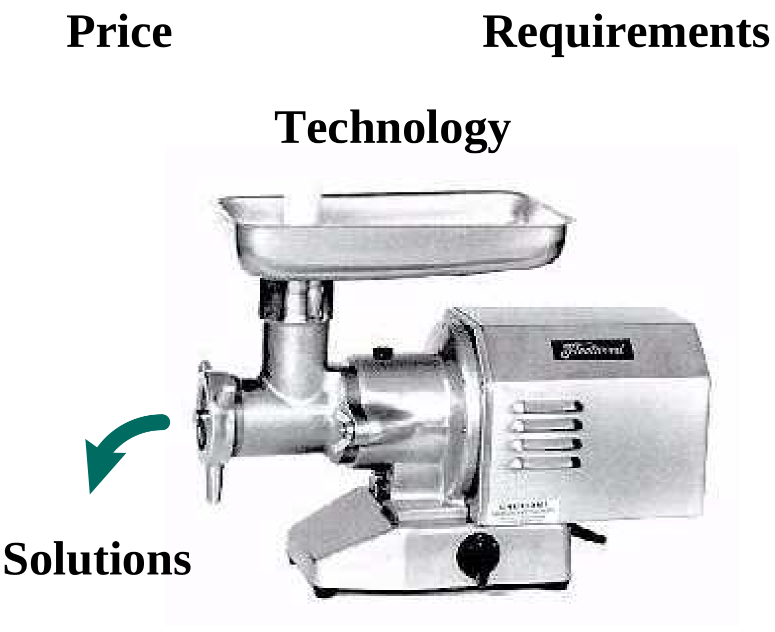
Run II Off-line Design

- Major Goals:
  - » 20 times more events, same processing and access time as Run I.
  - » Affordable, reliable, easy to use...
- Design Issues
  - » Disk Storage: What will the role of disk storage be?
  - » Serial Media: What kind(s) of serial media should be used?
  - » Robotic Storage: What will be the rolle of robotic storage?
  - » Storage vs. CPU: Can some storage be replaced by CPU?
  - » Processing CPU: What relative quantities of “farm” and “analysis” CPU will be needed?

Run II Off-line Design

- Design Issues (continued)
  - » Analysis CPU: What are the relative roles for Central server, Workgroup server and desktop analysis computing resources?
  - » Data to CPU Relationship: How tightly coupled should the analysis CPU and DATA be?
  - » Data Model: Granularity of access and clustering strategy; event, physics object or attribute.
  - » Data Access Model: Access patterns including off-site use and replication.

Choosing the Right Model



- Put price, technology and requirements in. Get the solutions out.

Estimated Prices for 1999

Area	Price Est.	Comment
Serial Media	$1-3/GB	
Robotic Storage	$100/slot	
Tape Drives	$5-120K each	
Disk Storage	$60/GB	
CPU - Reconst	$25/MIPS	PC farms may be lower
CPU - Analysis	$40/MIPS	

- Based on current prices with best guess estimates for price drops over next 2-3 years.

CPU Choices

- For us, there are two worlds of CPU: Distributed Network (DN) and Scalable Multi-Processor (SMP).
- Traditionally, DN has been used for straight forward, CPU intensive tasks like reconstruction, and SMP for less well-defined tasks, like analysis.
- We may be forced to leverage cheap DN CPU to do some traditional analysis-type projects.
- DN CPU may be slightly less convenient and have longer project setup times. Data access will be slower.

Disk and Robotic Storage

- Disk Storage - 7 TB/year
  - » Enough to keep thumbnail info (μDST) for each event.
  - » Temporary staging + analysis samples.
- Robotic Storage - 100 TB/year
  - » Latest (2-4 weeks) RAW data for reconstruction.
  - » Most heavily used Analysis samples.
  - » Some samples rotated in periodically from “the shelf”.

The Serial Media Labyrinth

- Although serial media, even in robotic storage, is still much cheaper than disk, the price has not decreased as rapidly as our data volumes have increased.
- In fact, relative to Run I (8mm video + data grade), the cost per GB has gone up.
- There are several related factors which need to be considered in choosing a technology:
  - » Drives: The cost for drives varies by almost 2 orders of magnitude. Head life and maintenance costs vary widely.
  - » Robotics: The cost of robotics scales with the number of slots, and is not decreasing. The number of slots per library must be balanced with mounting arm and drive contentions.
  - » Media: In some cases, lowest cost media requires expensive drives. Cartridge capacities are important to reduce slots.
  - » Software: Not all robotics are supported by all HSM software.

Possible Storage Options

- One media type: All data will be stored on the same serial media type which is robust, works in the robots but maybe expensive.
- Two media types: Infrequently needed data can be stored on less expensive “archival” media.
- Disposable data: As Distributed Network CPU continues to decrease in price, it may be cost effective to re-create, rather than store some processed information.

Three Models Illustration

Data Type	Size/event
Raw	250 kB
T1	300 kB
T2	50 kB
T3	5 kB

Media Type	Price	Effective Event Size for Model 1/ 2/ 3
M1	$2/GB	605 / 355 / 395 kB
M2	$1/GB	0 / 250 / 0 kB
M3 (virtual)	$25/MIPS	0 / 0 / 30 kB

- Model 1: All data on Media 1 (M1).
- Model 2: Raw data on Media 2 (M2).
- Model 3: All data on Media 1 (M1). But, only 20% of reconstructed data stored as T1, all data stored as T2. Then additional 10% remade, on demand, to T1 from RAW.

Three Models Result

Comparison	Model 1	Model 2	Model 3
Cost per year (600M events)	$726 k	$576 k	$600 k
What is latency for a T2 event Request?	1 Mount & seek	1 Mount & seek	1 Mount & seek
What is latency for a T1 event request?	1 Mount & seek	1 Mount & seek	1 Mount & seek + CPU
How long would it take to get a (non-random) sample of T2?	n Mounts	n Mounts	n Mounts
How long would it take to get a (non-random) sample of T1?	n Mounts	n Mounts	n Mounts + CPU

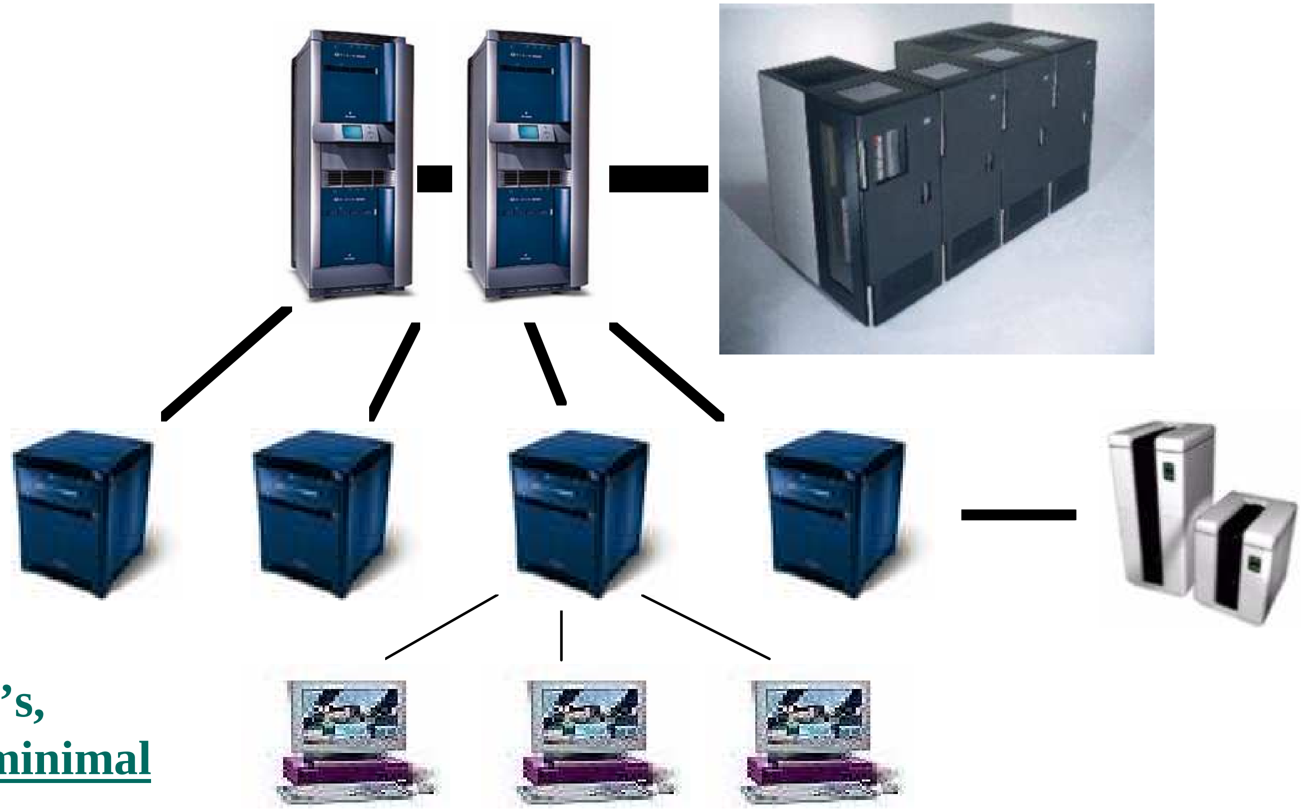
- Only media and CPU costs are included in estimates. Drives and other expenses may be significant.
- Model 2 is complicated by the need for additional drive varieties and data “imports” to robotics.
- Model 3 is complicated by software configuration. May be bottlenecks under heavy usage. Must be “transparent” !

Analysis Computing Model

Central SMP Servers + <u>most</u> Data Storage

Workgroup Servers + <u>some</u> Data Storage

Desktop X-terminals, PC's, workstations + <u>minimal</u> Data Storage



- Main Theme: Data and CPU tightly coupled.

Disclaimer: No specific solutions are intended.

Software

- Run Conditions Database Management
- Event Database Management
- Hierarchical Storage Management
- Process control for reconstruction and data serving
- Data Mining tools
- Plotting and data visualization tools

Conclusions

- We are looking forward to increasing our data statistics by 20 times in the upcoming run.
- We will employ our Run I experience to design data management systems for Run II.
- We will be hard pressed to supply enough storage and CPU to match the capabilities of the DAQ system, and the appetite for analysis.
- There are many technical decisions to be made in the next year. Which serial media technology? What is the role of robotic storage? How best can the various levels of CPU be exploited?