

Streaming the Data

Adam Lyon

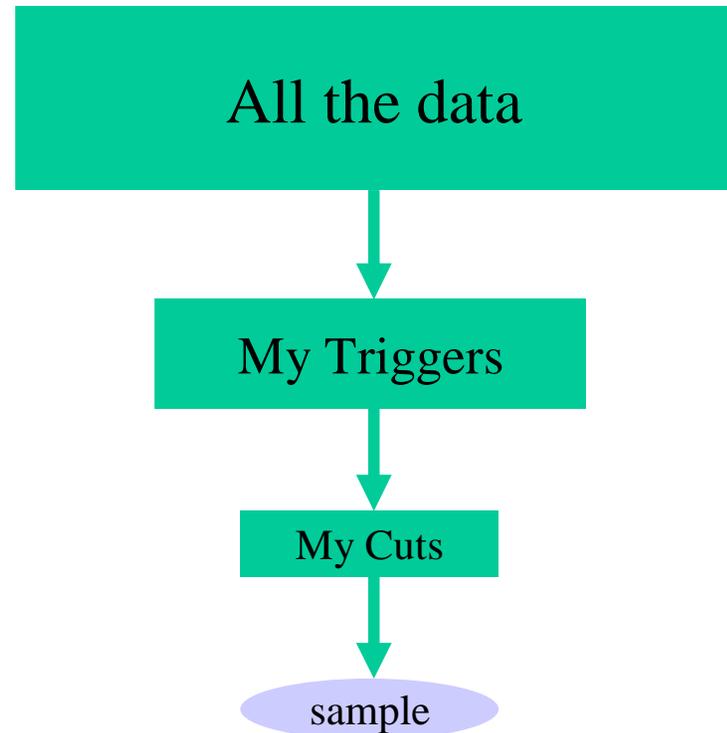
For the Analysis Tools Group

Oklahoma D0 Workshop Streaming Session

7/11/02

A “typical” analysis

- ◆ A Physicist does an analysis using events that pass certain triggers
 - ❖ A trigger is a set of (loose) criteria an event must pass before it is saved to tape
 - ❖ E.g. *EM_HIGH*: *Event has a high energy electron*
- ◆ Then make further requirements (cuts) on the data to look for signal
 - ❖ E.g. *electron > 20 GeV, not back to back with a jet, ...*

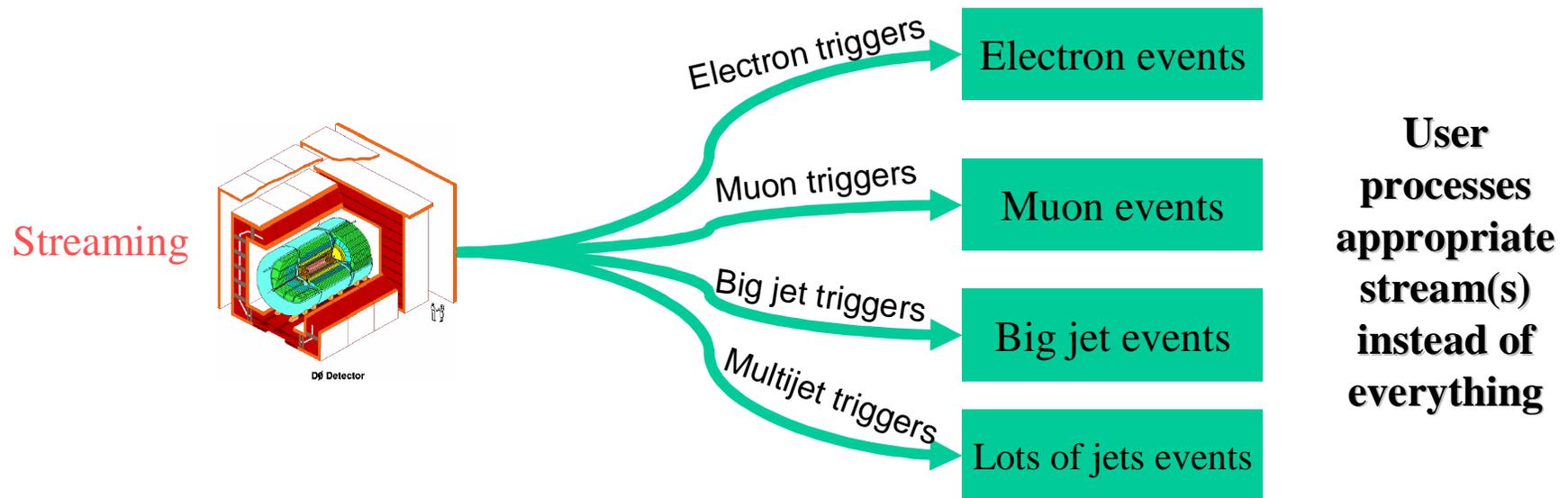
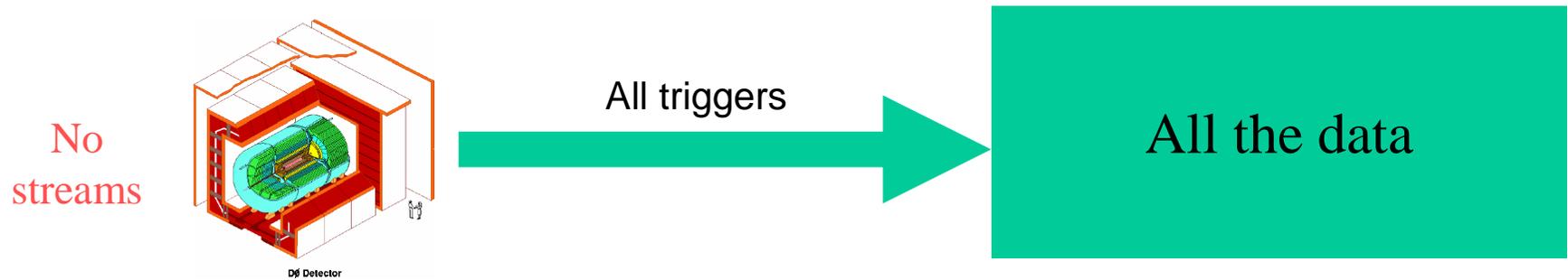


The problem

- ◆ All the data is huge!
 - ❖ My triggers are typically a very small subset of the data
 - ❖ **Can I avoid processing events I don't care about?**
- ◆ Running over the entire dataset is a big deal
 - ❖ Takes a long time
 - ❖ Unpleasant
 - ❖ **Extremely unpleasant if data comes from tapes**
- ◆ **Life is better if "All the data" → "Some of the data"**

Streaming – the ideal picture

- ◆ Separate data into streams



But... Streaming – The reality

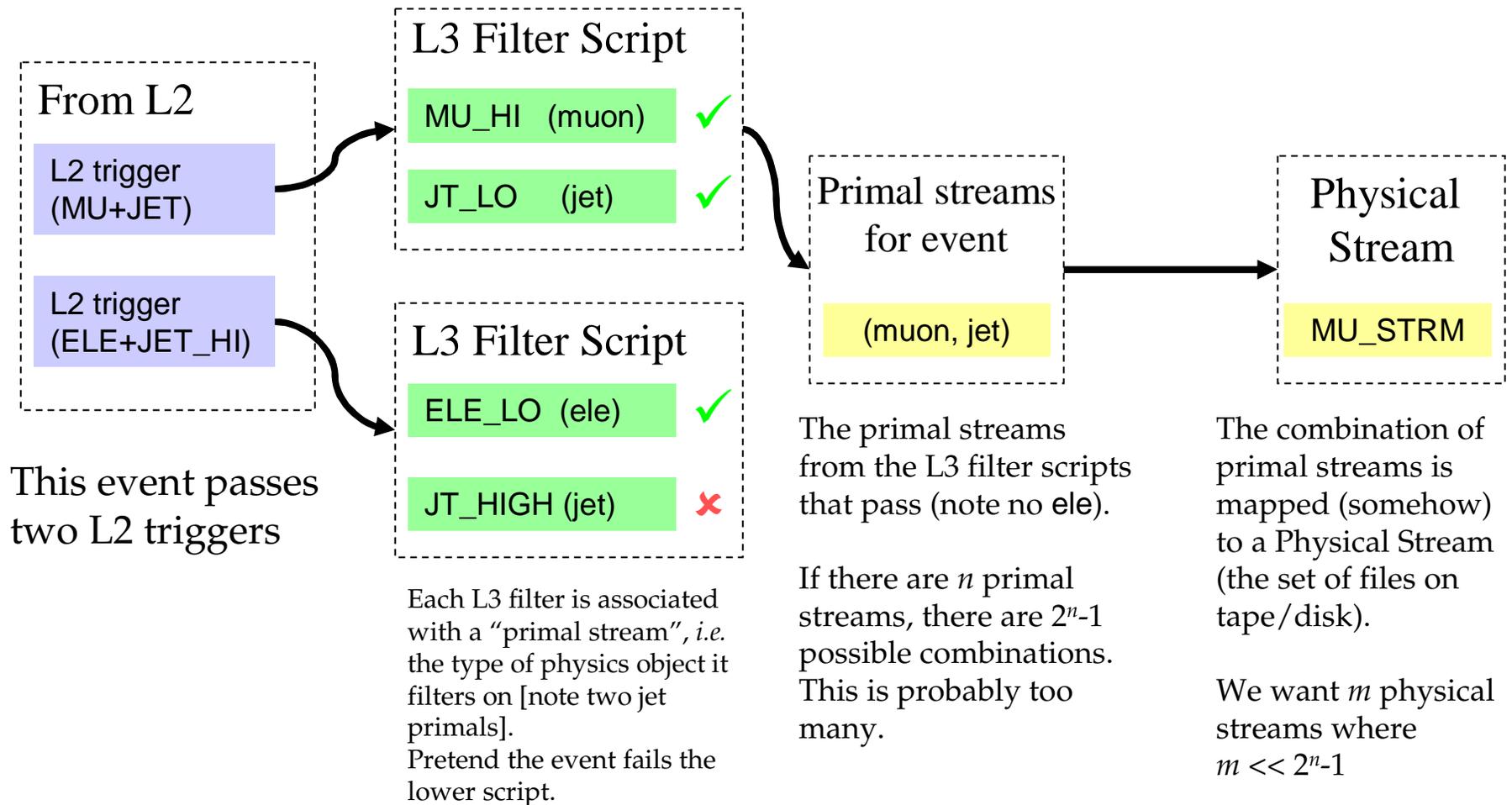
- ◆ Events may satisfy more than one trigger (*e.g.* electron and jet)
 - ❖ In Run I, such events were copied to >1 stream
 - ❖ Inclusive streaming
- ◆ Can't do that in Run II!
 - ❖ Many more events than Run I
 - ❖ **Tape costs are too high to have more than one copy of an event**
 - ❖ Online constraints make writing out copies of events difficult
- ◆ Run II: Exclusive streaming for regular physics data
 - ❖ Special monitor stream **will** be inclusive (for mark and pass)
 - ❖ Commissioning triggers will go to their own stream
 - ❖ Streaming will be done at L3 (**RAW data will be streamed**)

Exclusive streaming

◆ An event goes to one and only one stream

- ❖ Have some decision scheme when an event can satisfy more than one stream (more on this later)
- ❖ The decision is based on Level 3 physics objects (primal streams) for passing triggers (see next slide for diagram)
 - L3 physics objects are more stable than triggers – don't have to rethink streaming for every small change in a trigger list
 - L3 physics objects make it easy to have like events in the same stream
 - **BUT:** Will be able to override stream decision for specific triggers – e.g. Trigger X **always** goes to stream S (meant to handle commissioning triggers, but may be useful elsewhere)

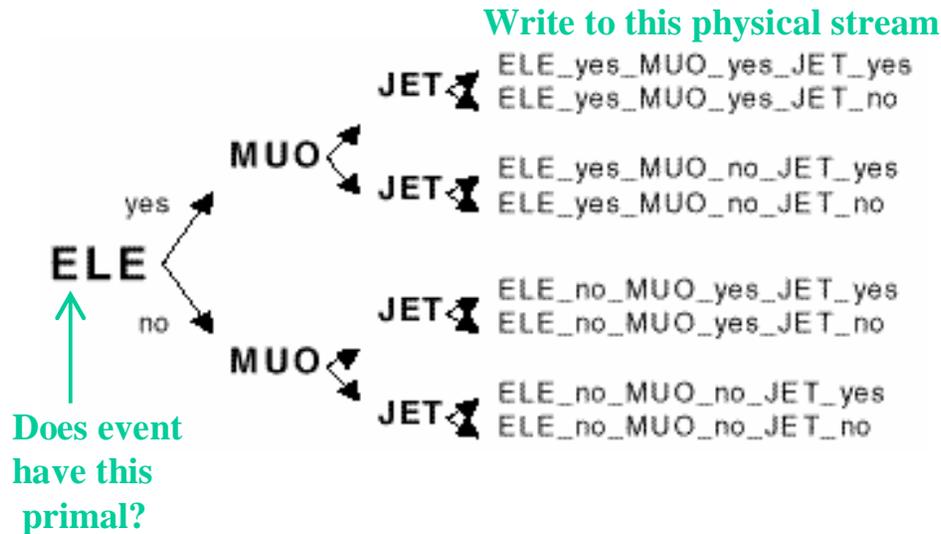
The Streaming Process: An example



Streaming Scheme Examples

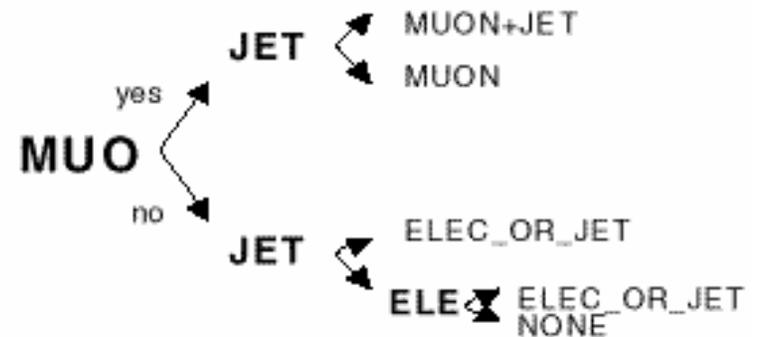
[from Jon Hays' document]

◆ Simple one-to-one:



- ❖ $2^n - 1$ physical streams
- ❖ Always n decisions
- ❖ No ambiguity

◆ Priority:



- ❖ Here, event with a muon and an electron goes to the muon stream
- ❖ Can sometimes decide in $< n$ decisions

Cannot avoid possibility that events from the same trigger will go to more than one stream!

Processing Data

- ◆ Processing data is more complicated
 - Events for a trigger may be in more than one stream
- ◆ Analysis Tools Group Plan –
Streaming is transparent to the user
 - ❖ User specifies trigger(s) to analyze
 - ❖ Some tool (that the ATG writes) figures out the needed streams and generates a project for SAM
 - ❖ User never knows about streams
 - ❖ Tools will be available to calculate luminosity

Are we ready to stream?

- ◆ Much of the infrastructure for streaming is already in place
- ◆ We need tools to be written to make data analysis and luminosity calculation easy (this is the purpose of the Analysis Tools Group)

How do we decide how to stream?

- ◆ The physics groups and the trigger board must determine the Streaming Scheme (not the ATG)
 - ❖ Stream scheme is meant to be flexible and can be altered to accommodate new triggers and new trigger systems. **But trigger scheme should be stable for long periods!**
 - ❖ What primal streams do we want (*e.g.* just jet or jet_hi and jet_lo)?
 - ❖ How do we map the primal streams to the physical streams?
 - ❖ **The ATG is writing a tool to simulate streaming that you can use to test ideas. More on this later.**

How to stream?

- ◆ The stream scheme will be a question of priorities:
 - ❖ With exclusive streaming, impossible to please everybody!
 - Streaming is designed to streamline data access for select analyses
 - Hopefully, many analyses will see some benefit
 - Some analyses (with many final states) will still look at all of the data
 - ❖ Which analyses will be the “select”?

Schedule

- ◆ Soon after OK workshop –
Physics groups and trigger board starts experimenting with streaming schemes using simulation tool
- ◆ By end of August, 2002 –
Streaming is tested online using a strawman scheme from Greg Landsburg
(read <http://hep.brown.edu/users/Greg/streaming/st.htm>)
- ◆ Turn on from shutdown (10/2002) – Streaming is online and operational (need streaming scheme by this time)

Streaming Summary

- ◆ An analysis based on specific triggers should (hopefully) not have to process all of the data
- ◆ Streaming should be transparent to the users
 - ❖ Analysis user only worries about triggers
- ◆ Streaming should be flexible, but stable
 - ❖ Easy to alter by authorized experts
 - ❖ Built-in sanity checks to avoid online problems
- ◆ Streaming should be beneficial to most users
 - ❖ Exclusive streaming won't please everyone, but it should help many, even if only slightly
- ◆ You can try it out! -- Stream Evaluation Tool

We need you!

- ◆ The streaming scheme will be difficult to determine (can't please everybody all of the time). Your input is important.
- ◆ Are there special streaming requirements not covered here? Let us know!
- ◆ Come to the ATG meetings
Tuesday's 1pm Racetrack (WH7X)
d0-atg@fnal.gov