



# CMS Data-to-Surface

K. Sumorok

MIT

US\_CMS DOE/NSF Review

FNAL, June 5, 2002

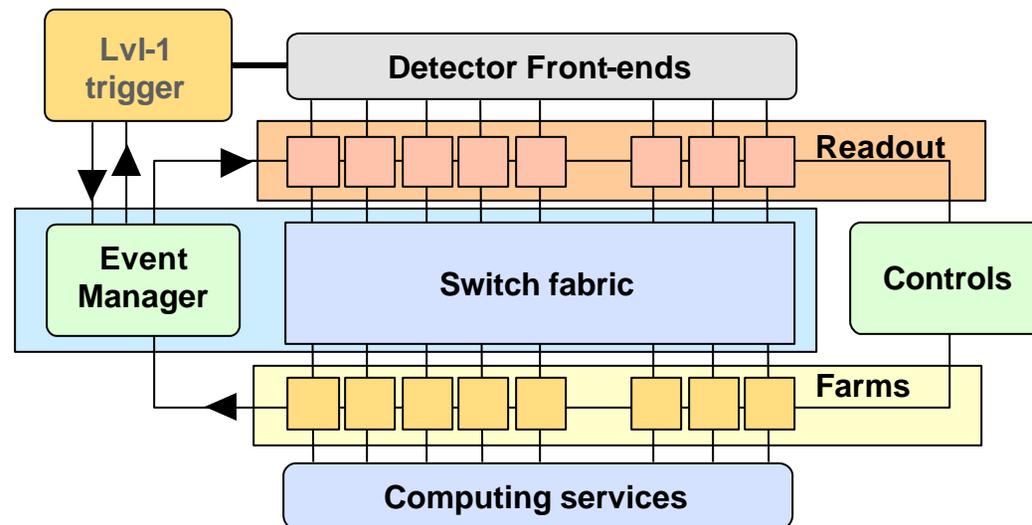
- Architecture, basic review
- Readout
  - ◆ Previous Work
  - ◆ Data-to-Surface
  - ◆ Current Work
- Plan of Work/Summary

# **Architecture, basic review**



# CMS DAQ Architecture

## ■ Current DAQ elements

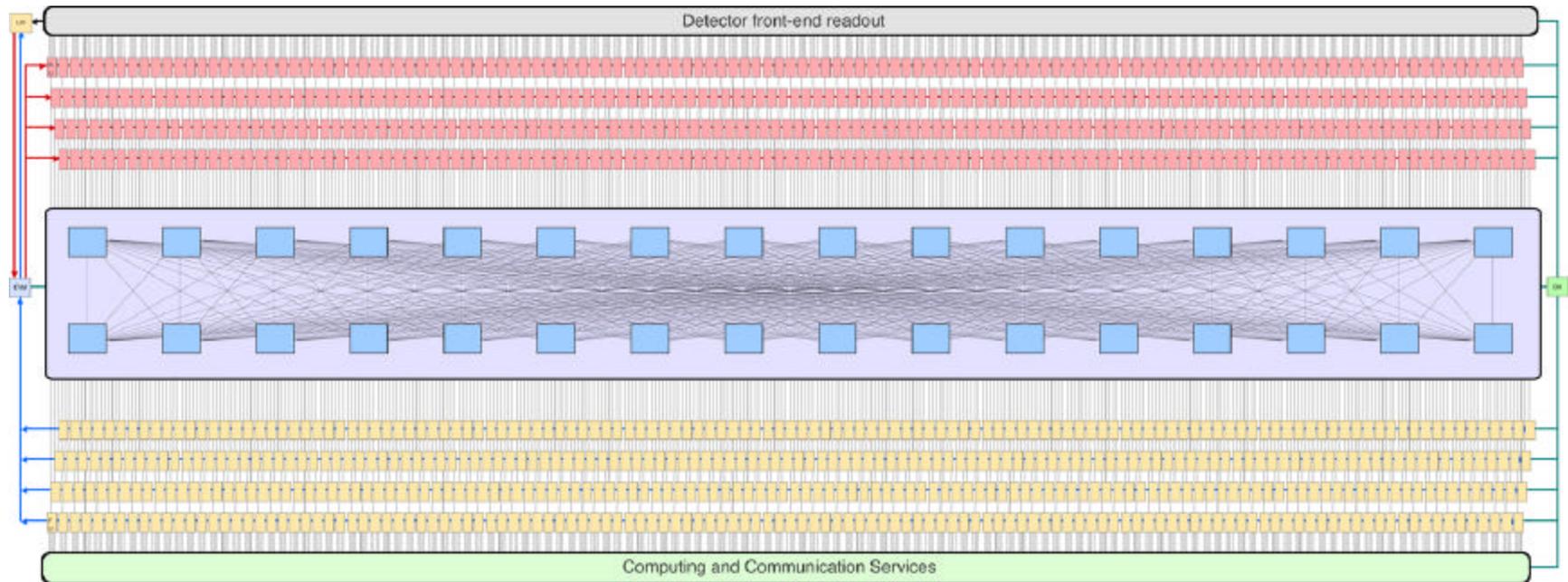


- Readout (units/drivers/buffers/...)
- Switching network(s)
- Processor Farm(s)
- Control & Monitor System



# CMS DAQ Design: old

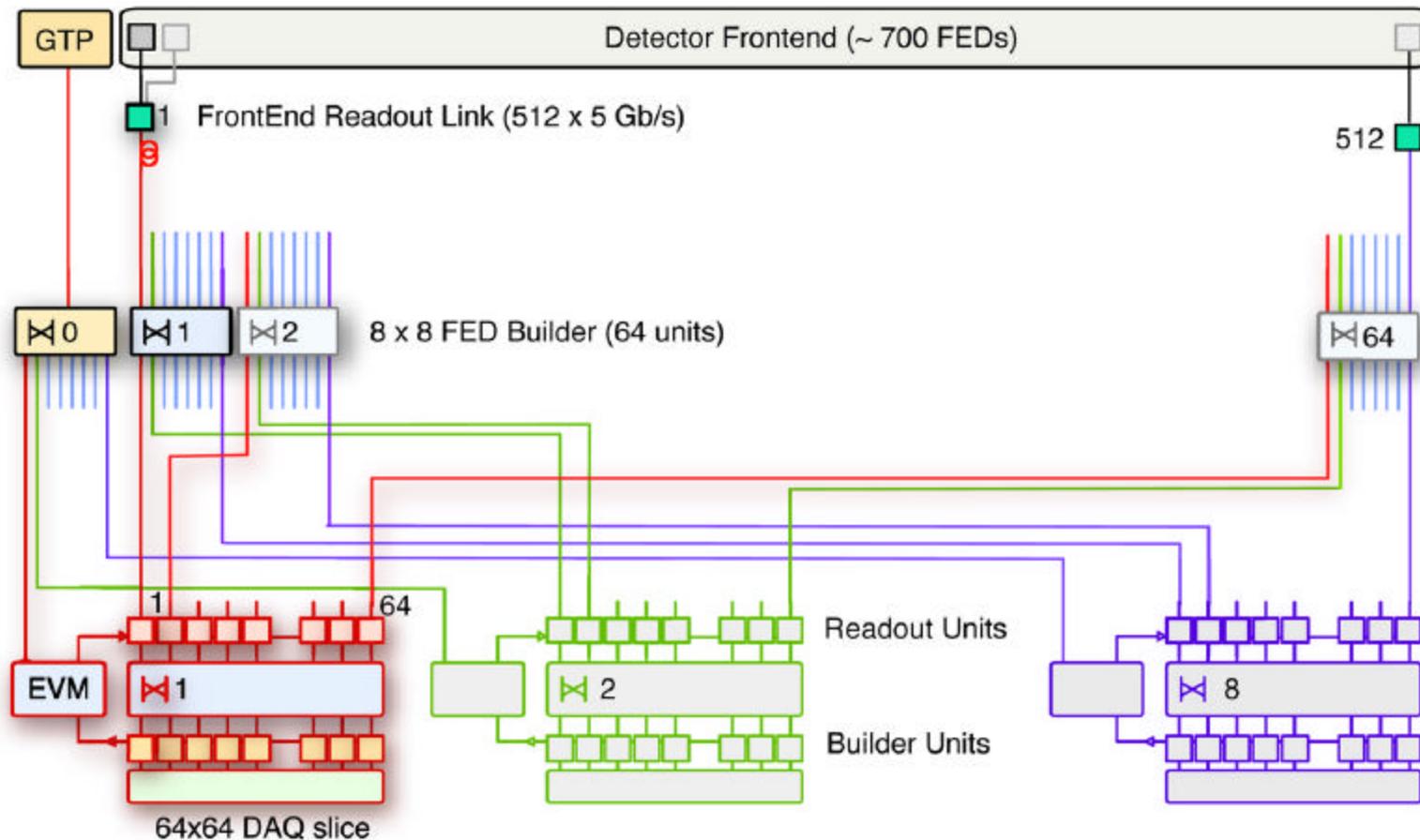
- Possible implementation with 2x16 switches of 32x32
  - ◆ Two-stage network with no intermediate buffering
  - ◆ Must function at near full capacity, at 100 kHz





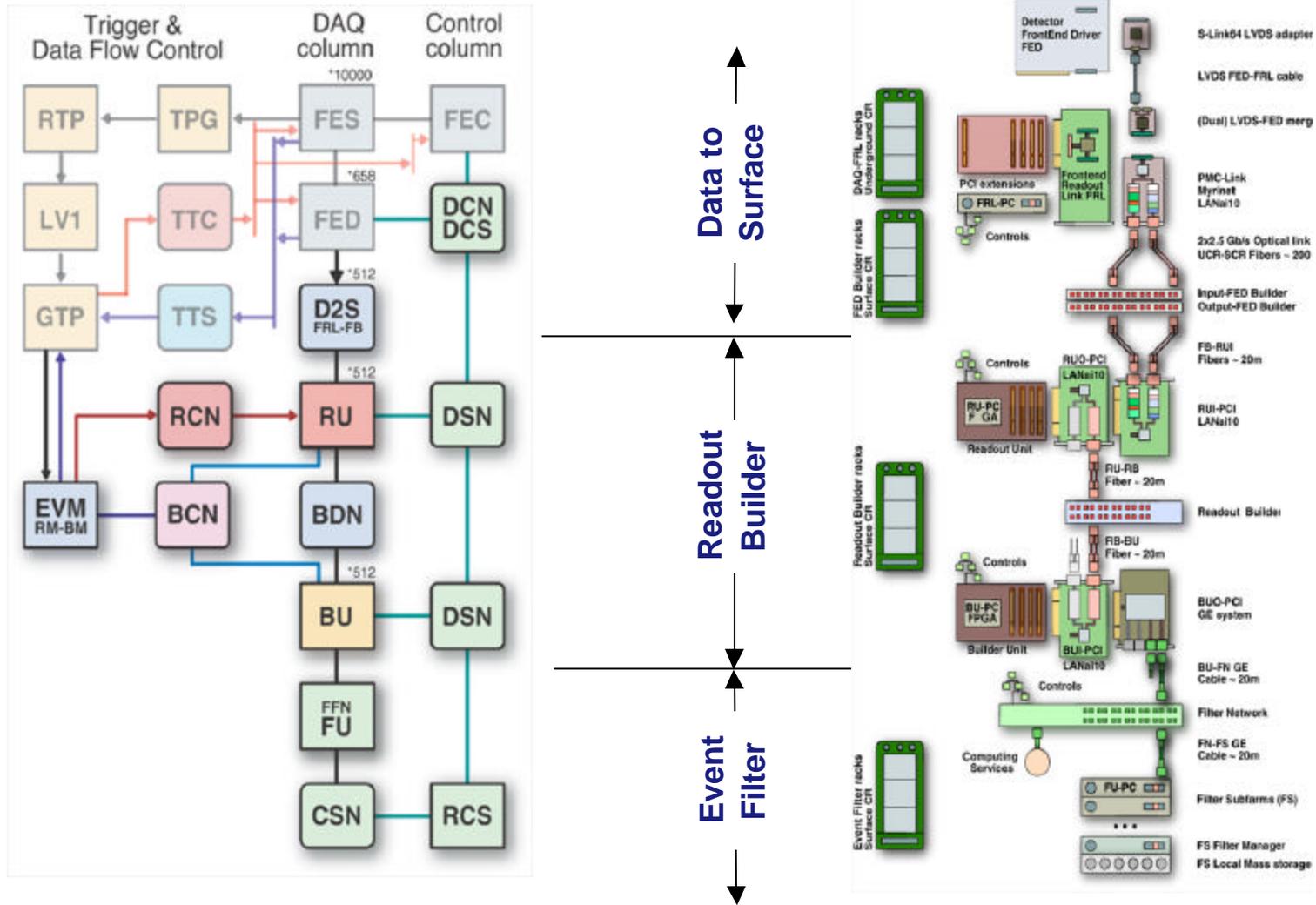
# CMS DAQ design: new

- Two-stage network with intermediate buffering





# CMS DAQ: elements (in a “column”)





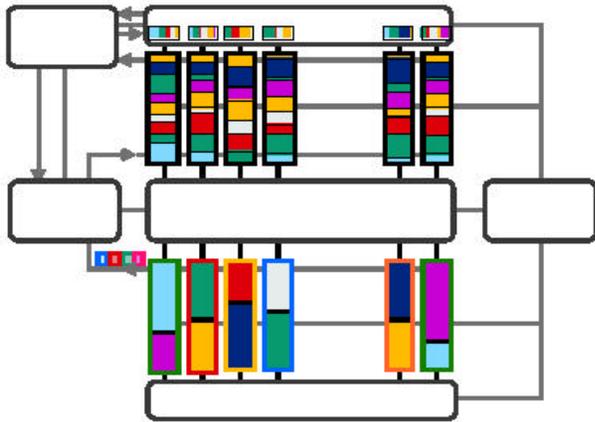
## New design

- Introducing buffers between the FED and the Readout Units decouples the two (EVB) stages
  - ◆ Data-to-Surface (D2S) system operates at 100 kHz with ~2kB fragments
  - ◆ Readout Builder (RB) system operates at 12.5 kHz with ~16 kB fragments
  - ◆ As a result: both systems can be implemented today (i.e. to design uncertainty at this point)
    - All that remains: final choice of technology (and price drops)

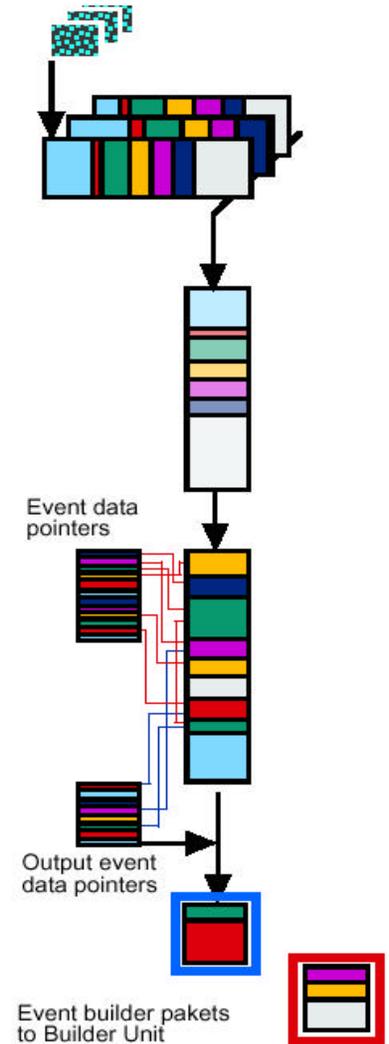
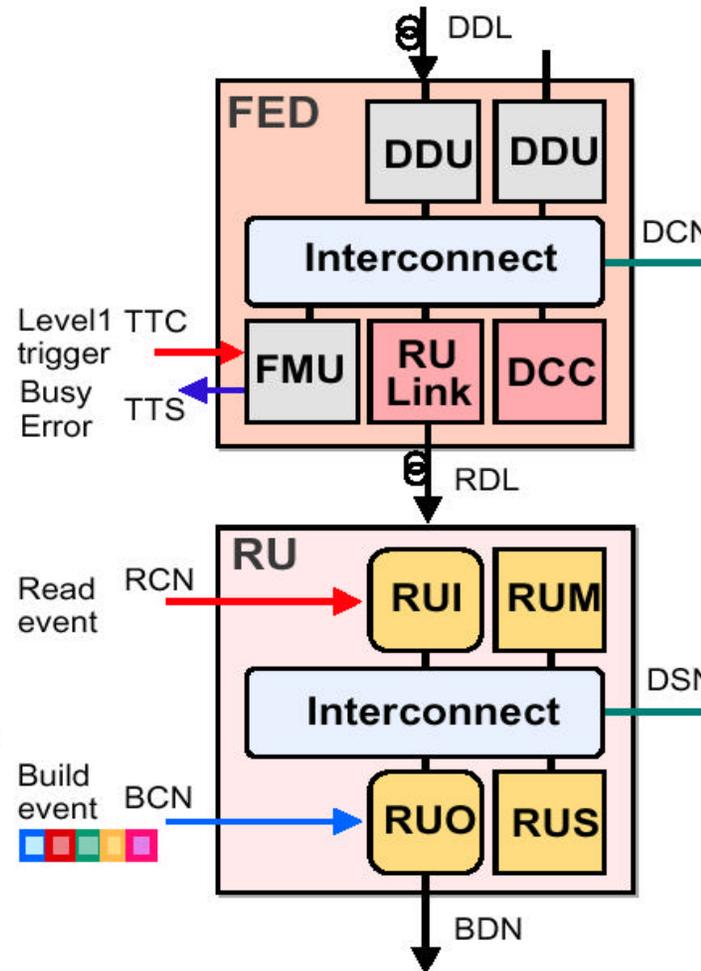
# **Readout: previous work**



# FED-RU data flow



Level-1 trigger is distributed to all DDUs by the TTC system. Each DDU reads, (digitizes), formats, checks and buffers the frontend data in less than 10 $\mu$ s. When RU-Link ready, buffered (up to 8) event fragments, are sent to the FED output (RU-Link) in the same order of generation. RU-link system behaving as a FIFO, RU reads a pending event on response from a event manager command. RU buffer half full and data time tags are monitored and signals are sent via TTS to the GTP for data flow controls.



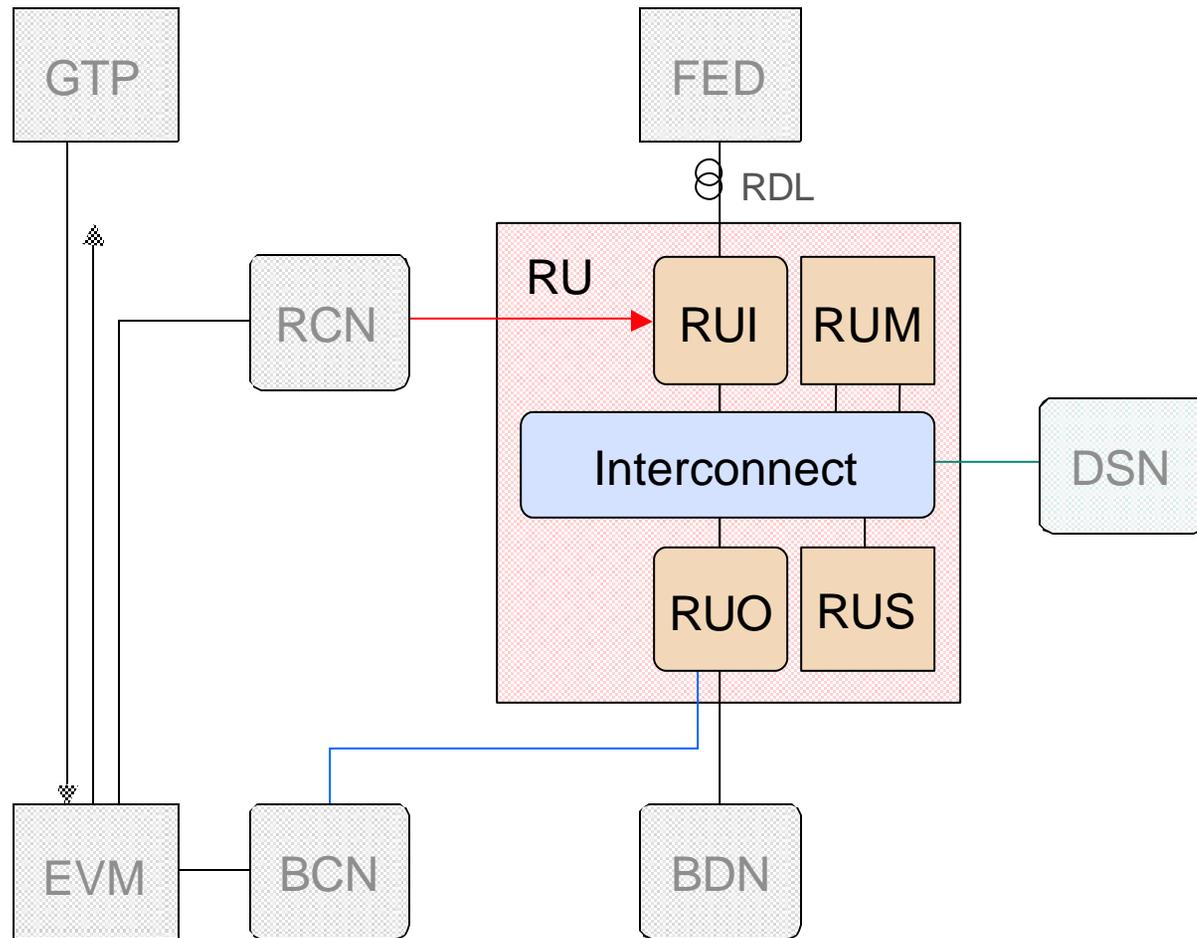


# Readout Units

- **Currently, dual-ported data access**
  - ◆ Additional ports for control
  - ◆ DAQ element with lowest latency ( $\sim\mu\text{s}$ ), highest rate
  - ◆ Basic tasks:
    - Merge data from N front-ends (detector-dependent units: DDUs)
    - Send data onto processor farm
    - Store the data until no longer needed (data sent or event rejected)
  - ◆ Issues:
    - Input interconnect (bus/point-to-point link/switch)
    - Output interconnect (bus/point-to-point link/switch)
    - Sustained bandwidth requirement (200-800 MB/s)



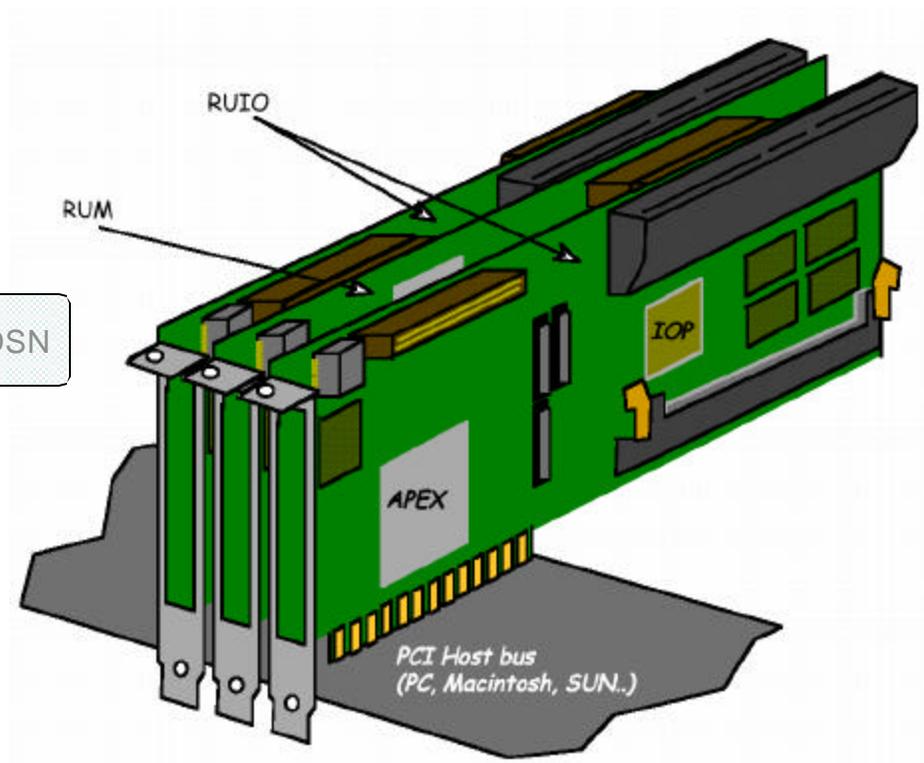
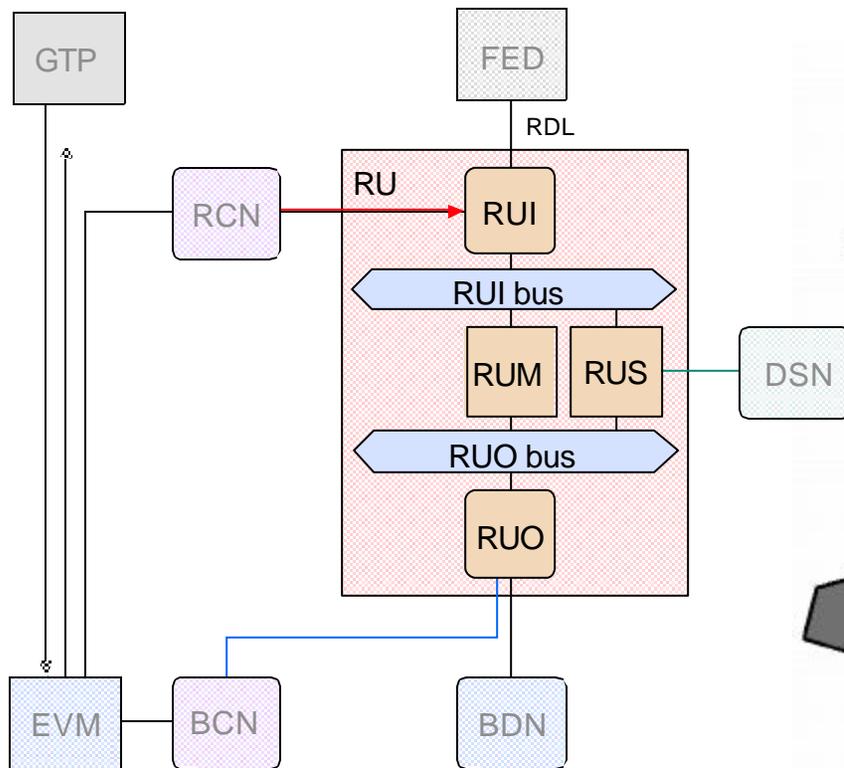
# Readout Unit Architecture





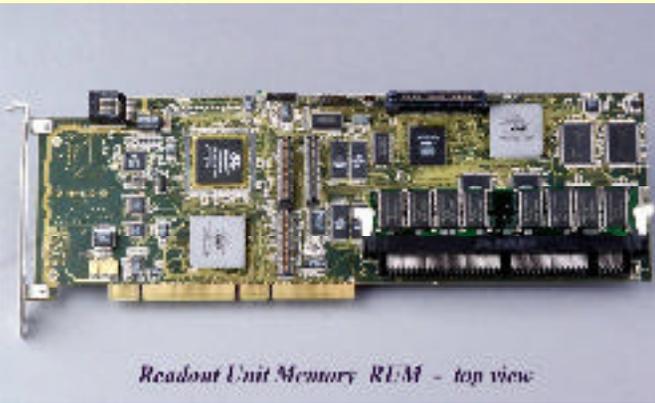
# Readout Unit – P2

- Evolution of RDPM developments (1994-98)





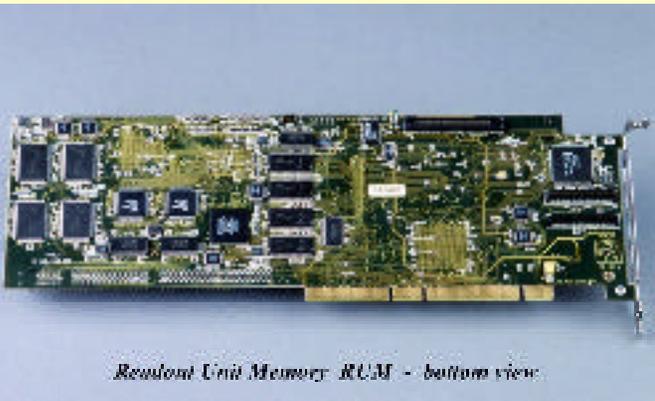
# RU-P2 elements



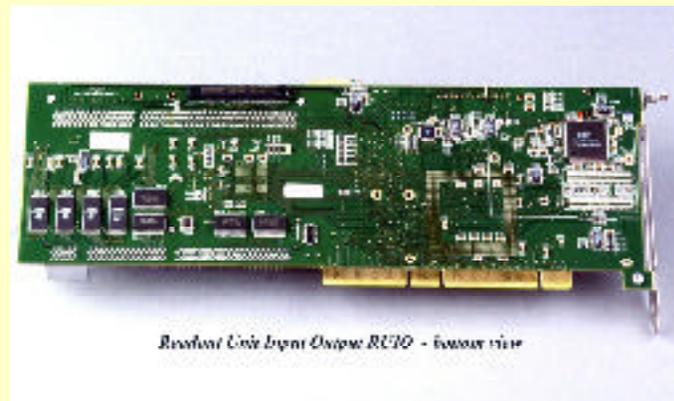
*Readout Unit Memory RUM - top view*



*Readout Unit Input Output RUIO - top view*



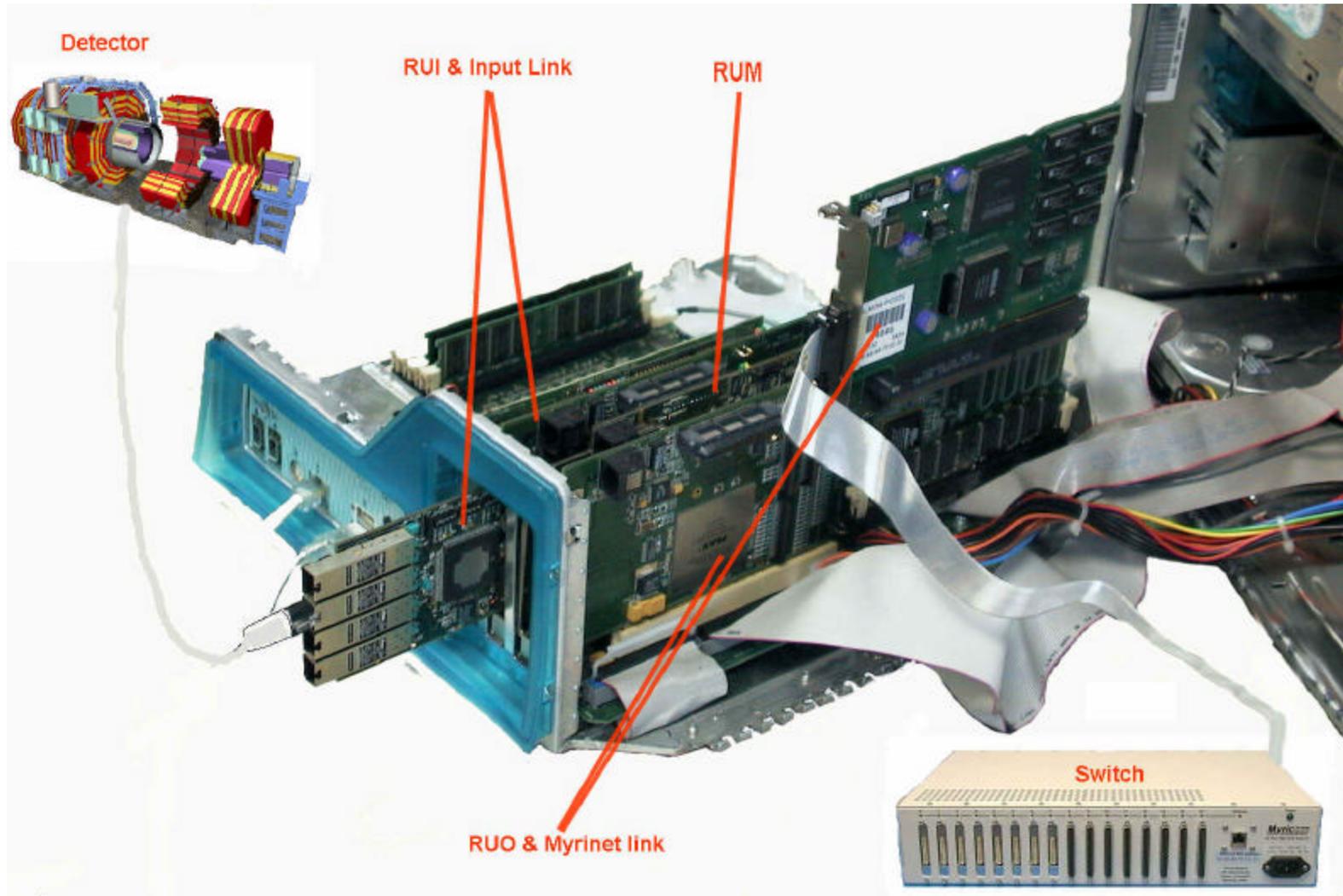
*Readout Unit Memory RUM - bottom view*



*Readout Unit Input Output RUIO - bottom view*



# RU-P2 teststand



**Readout: data to  
surface (D2S)**



# Data-to-Surface system (I)

## ■ Purpose:

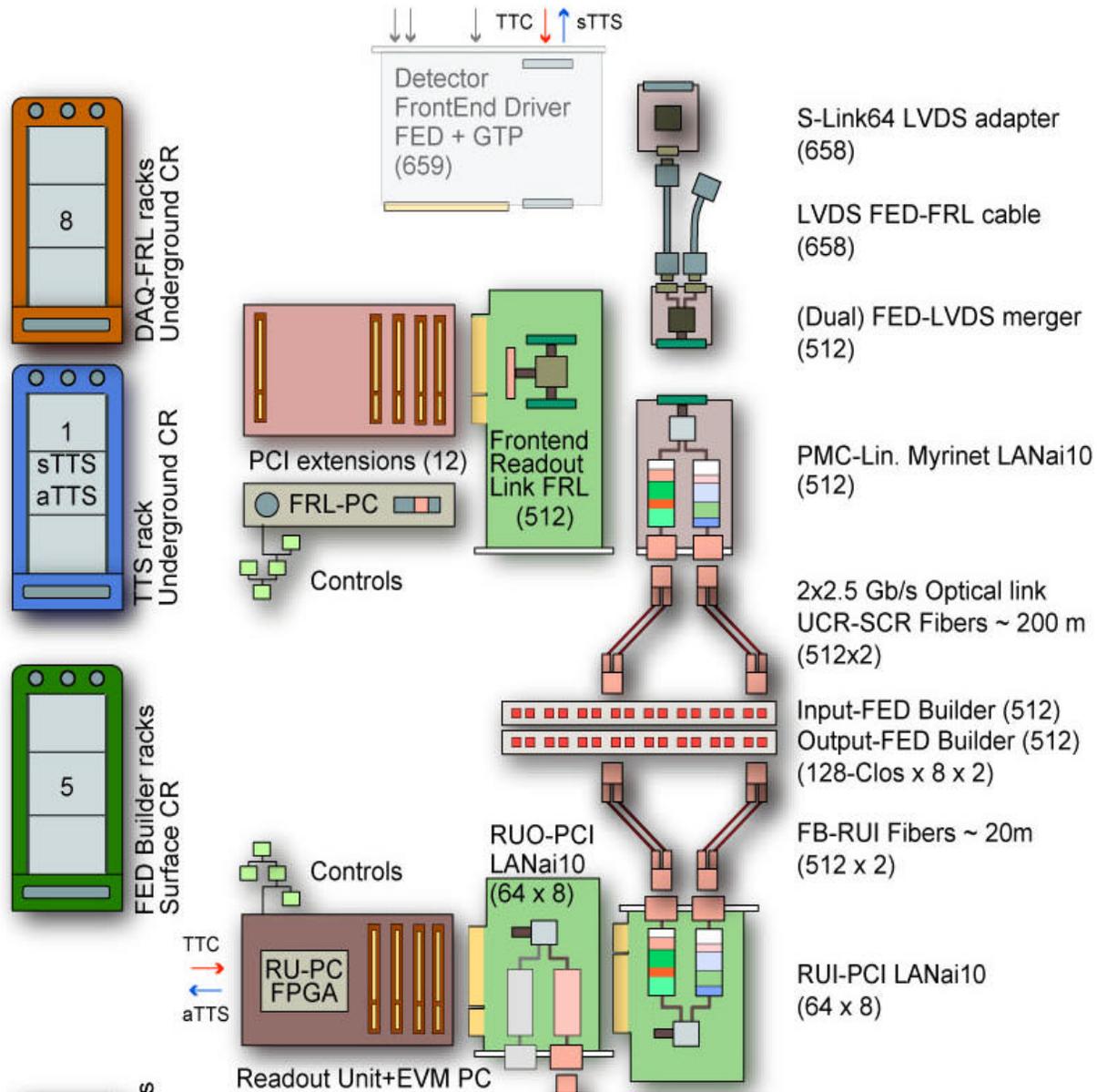
- ◆ Transport data from the detector hall to the counting room (~200m)
- ◆ Offload the FEDs (which have limited memory) as quickly as possible
- ◆ Merge data fragments from the FEDs to roughly equal-size “super-fragments” for the Readout Units
  - To increase efficiency of Readout Builder
- ◆ Route events to the (available) Readout Builders

## ■ Data flow:

- ◆ FED (via SLink64) to FRL; FRL merges up to 2 FEDs
- ◆ FRLs (via 8x8 switch) to Readout Unit (its Input)



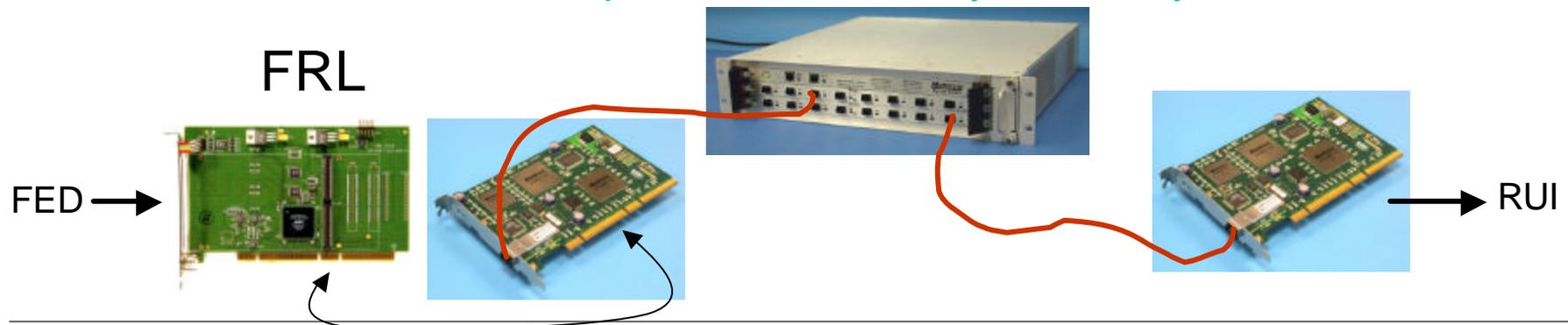
# Data-to-Surface (II)





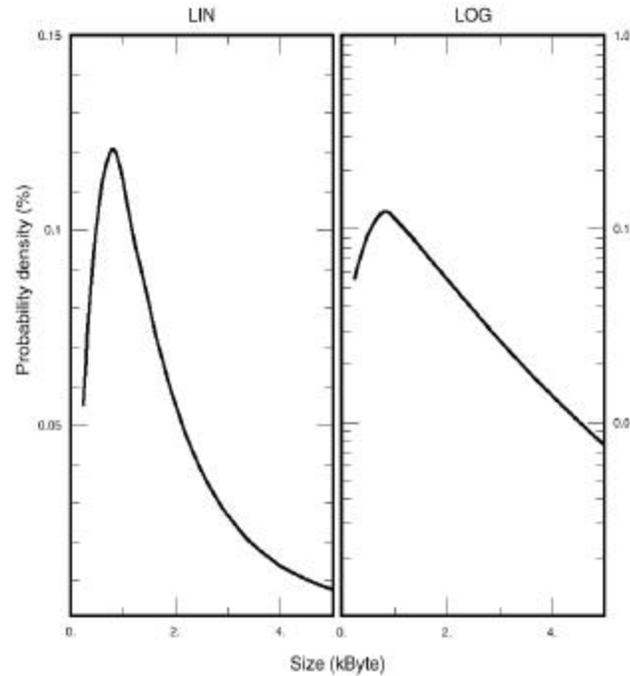
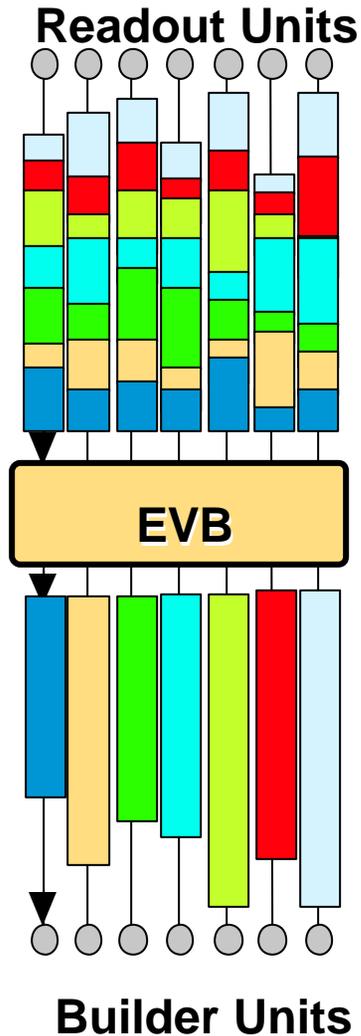
## Data-to-Surface (III)

- Requirement: sustain 200 MB/s throughput/link
  - ◆ For a switch with ~8-16 ports, need either traffic shaping, or higher bandwidth (efficiency is ~50%, see next transparency)
    - Opt for 400 MB/s throughput (assume little intelligence before the Readout Unit)
    - Wait for Lanai10 from Myrinet (2x2.5 Gb/s links)
    - And since these links are fiber links, the data can be transported over the 200m required – on the way to the Myrinet switch





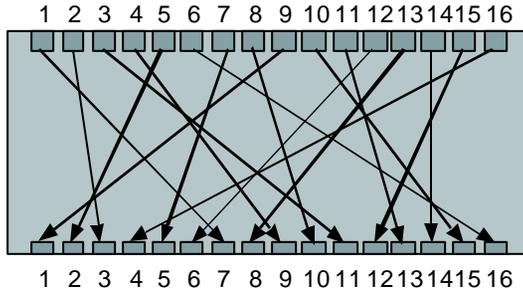
# FED Builder Inputs



Log-normal distribution  
example: Average = 2 kB, RMS = 2 kB  
mimics CMS data readout



# Myrinet: understood up to 32 ports



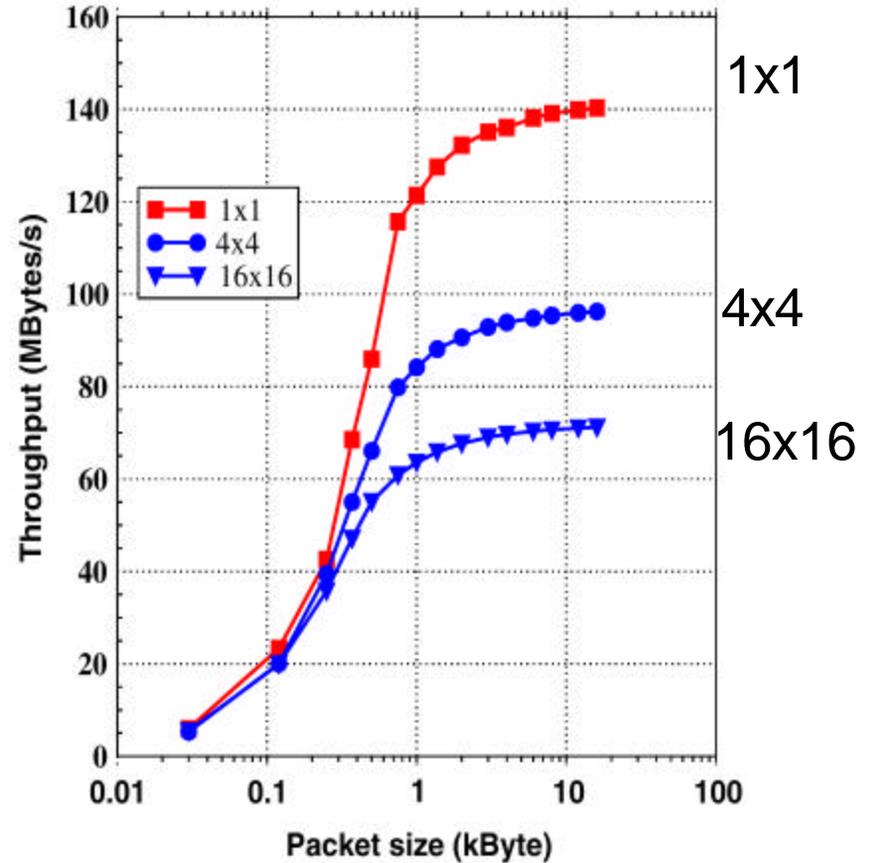
sources send, independently,  
to a **random destination**  
according to a uniform  
distribution

Efficiency:

4x4: 69 %    expect 68%  
16x16: 51 %

limited by head-of-line blocking

### Point to Point NxN Random Destinations



measured at destinations

# **Readout: program of work**



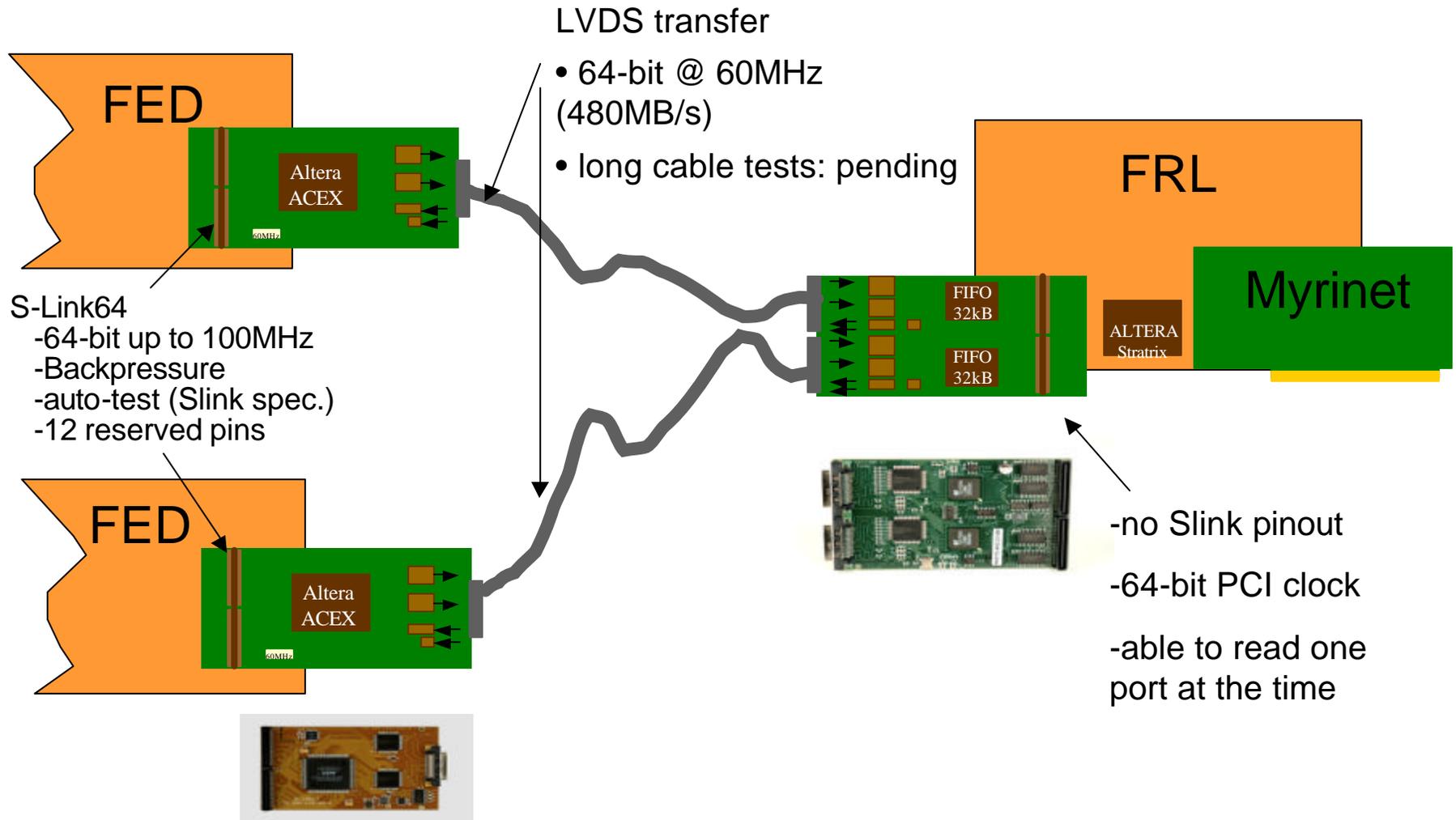
## D2S: ongoing/future work

### ■ Pending items:

- ◆ Merging of FEDs into the FRL
  - Currently being addressed
- ◆ Event Building with random traffic (no Event manager, no traffic shaping) on the FED Builder
  - Has been done with the old Myrinet hardware
    - ⇒ Will repeat exercise whenever Lanai10 arrives (Q3/Q4 of 2002)
- ◆ Synchronization with Trigger and Readout Builders
  - Working on Global Trigger emulator, FED emulator, etc.

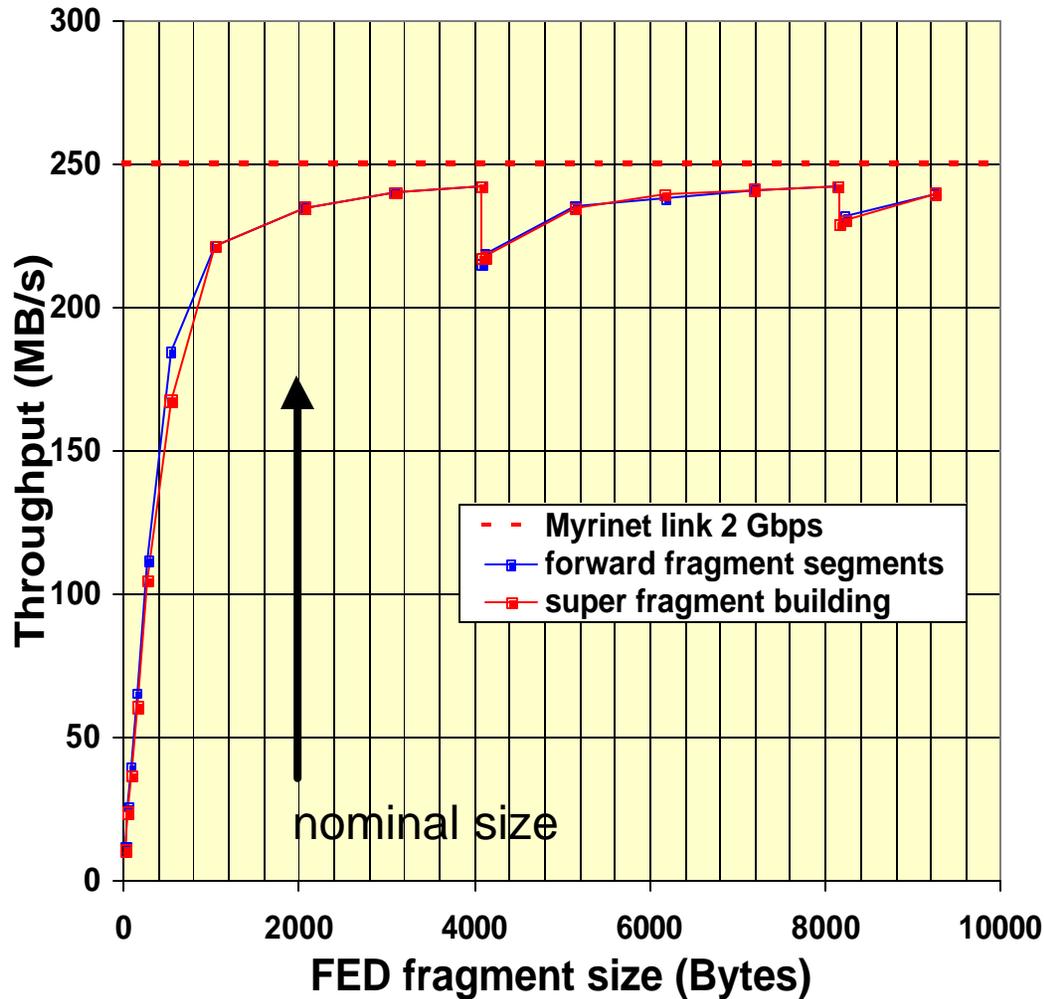


# FED Readout Link (FRL)





# FRL: s-fragment building performance



## NIC Input:

- for 2 kByte fragments
- throughput 230 MB/s
- or 115 kHz trigger rate

## NICOutput = RU Input:

- 16 kByte super-fragments
- 14 kHz rate



## Summary/Conclusion

- D2S: transport data to counting room, perform first-stage of event-building (16 kB subevents), multiplex events to Readout Builders
  - ◆ Should be in place for 2004-2005 detector readouts
  - ◆ Myrinet-based system works today
  - ◆ Want to use 2x2.5 Gb/s links
- Previous work on Readout Unit: RUM ok, currently in use in column; GIII card used for FRL prototype
- Longer-term: integration with other DAQ elements