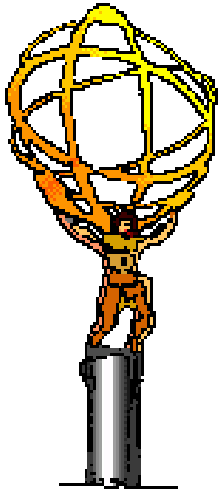
UNIVERSITÄT BERN

Performance of the final Event Builder for the ATLAS Experiment



HP Beck

LHEP – University of Bern

On behalf of ATLAS TDAQ DataFlow

15th IEEE NPSS Real Time Conference 2007

Fermilab, Batavia IL, 60510

April 29 – May 4, 2007

LABORATORIUM FÜR HOCHENERGIEPHYSIK
LHEP
UNIVERSITÄT BERN

ATLAS TDAQ DataFlow

H.P. Beck[1], M. Abolins[2], A. Battaglia[1], R. Blair[3], A. Bogaerts[4], M. Bosman[5], M. Ciobotaru[6], R. Cranfield[7], G. Crone[8], J. Dawson[3], R. Dobinson[4†], M. Dobson[4], A. Dos Anjos[9], G. Drake[3], Y. Ermoline[2], R. Ferrari[10], M.L. Ferrer[11], D. Francis[4], S. Gadomski[1], S. Gameiro[4], B. Gorini[4], B. Green[12], W. Haberichter[3], C. Häberli[1], R. Hauser[2], C. Hinkelbein[13], R. Hughes-Jones[14], M. Joos[4], G. Kieft[15], K. Kordas[1], A. Kugel[13], L. Leahu[16], G. Lehmann[4], B. Martin[4], L. Mapelli[4], C. Meessen[17], C. Meirosu[15], A. Misiejuk[12], G. Mornacchi[4], M. Müller[13], Y. Nagasaka[18], A. Negri[6], E. Pasqualucci[19,20], T. Pauly[4], J. Petersen[4], B. Pope[2], J. Schlereth[3], R. Spiwoks[4], S. Stancu[6], J. Strong[12†], S. Sushkov[5], T. Szymocha[21], L. Tremblet[4], G. Unel[4,6], W. Vandelli[4], J. Vermeulen[15], P. Werner[4], S. Wheeler-Ellis[6], F. Wickens[8], W. Wiedenmann[9], M. Yu[13], Y. Yasu[22], J. Zhang[3], H. Zobernig[9]

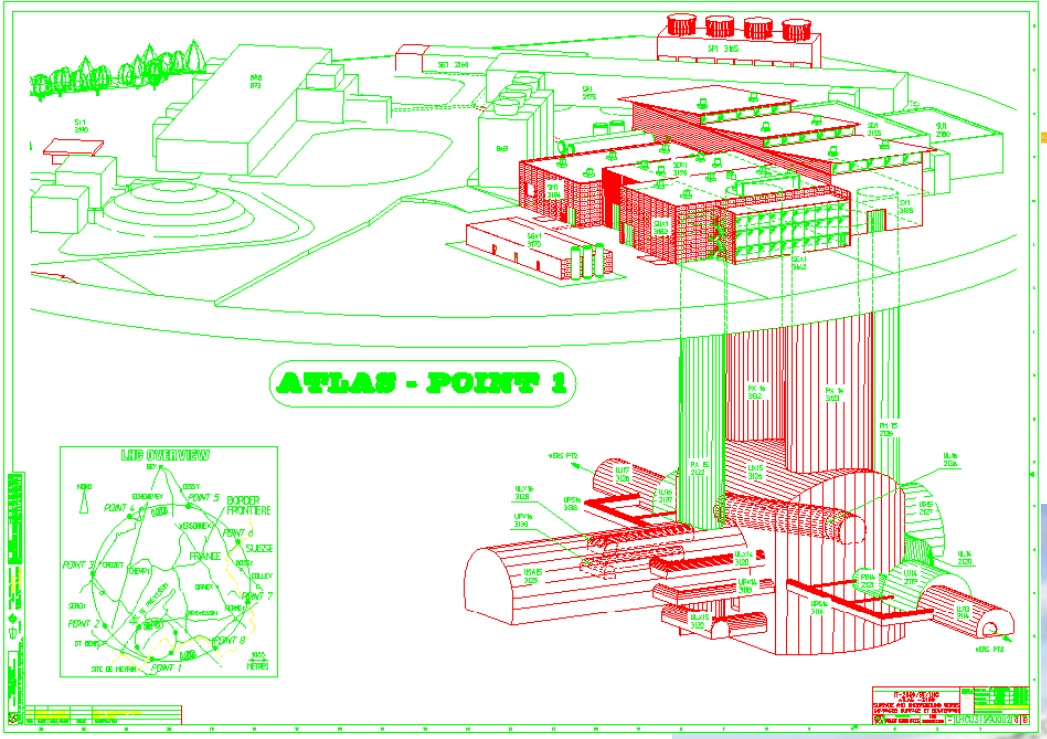
† deceased

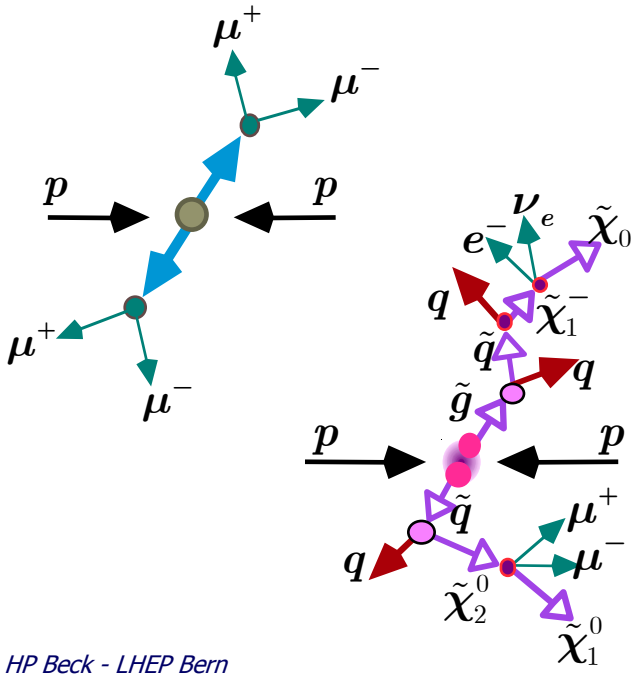
1. Universität Bern, Switzerland
2. Michigan State University, Ann Arbor, MI
3. Argonne National Laboratory
4. CERN, Geneva, Switzerland
5. Inst. de Fisica de Altas Energias (IFAE), Universidad Autonoma de Barcelona, Spain
6. University of California, Irvine, CA, US
7. University College, London, UK
8. CCLRC Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, UK
9. Univ. of Wisconsin, Madison, WI, US
10. INFN Sezione di Pavia, Italy
11. Laboratori Nazionali di Frascati, Italy
12. Physics Department, Royal Holloway College, University of London, Italy
13. Universität Mannheim, Germany
14. University of Manchester, UK
15. NIKHEF, Amsterdam, The Netherlands
16. National Institute for Physics and Nuclear Engineering "Horia Hulubei", NIPNE-HH, Bucarest, Romania
17. CPPM Marseille, France
18. Hiroshima Institute of Technology, Japan
19. Universita di Roma "La Sapienza", Rome, Italy
20. INFN Roma, Rome, Italy
21. Henryk Niewodniczanski Inst. Nucl. Physics, Cracow, Poland
22. High Energy Accelerator Research Organization (KEK), Tsukuba, Japan

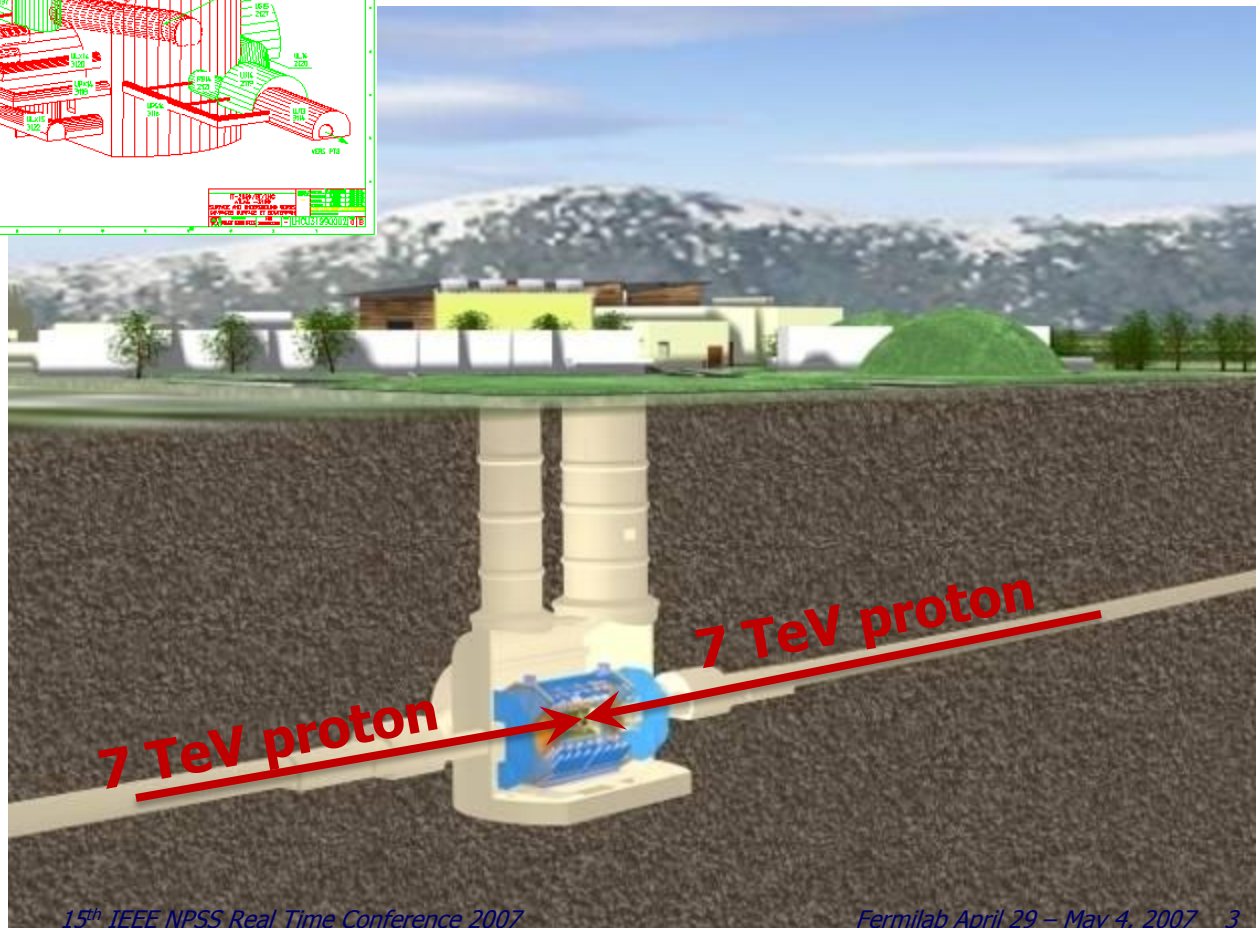
The ATLAS Experiment

Proton-Proton collisions at 7 + 7 TeV cms

40 MHz Beam crossing rate
~1 GHz pp-collisions



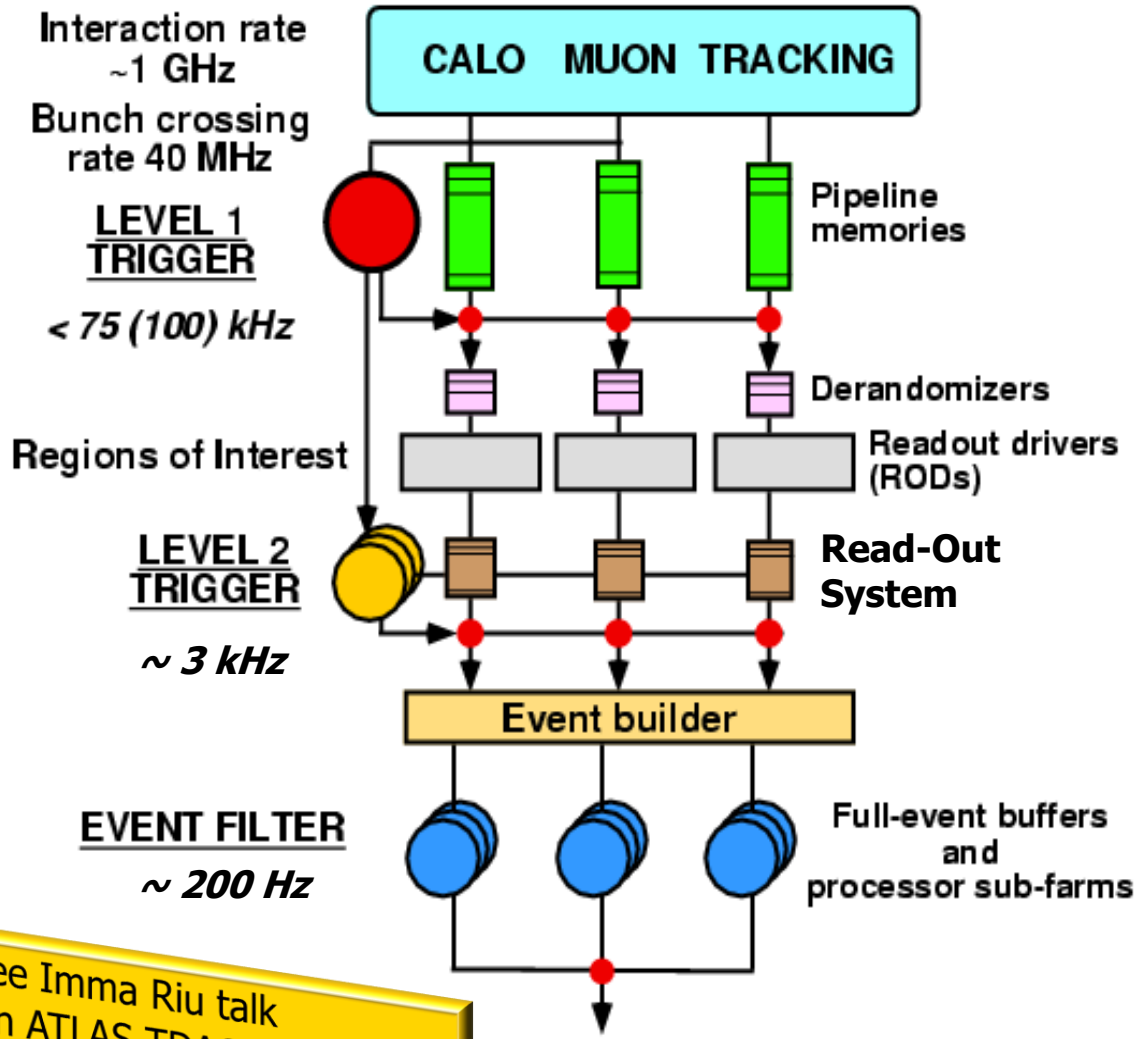




Three Trigger-Levels

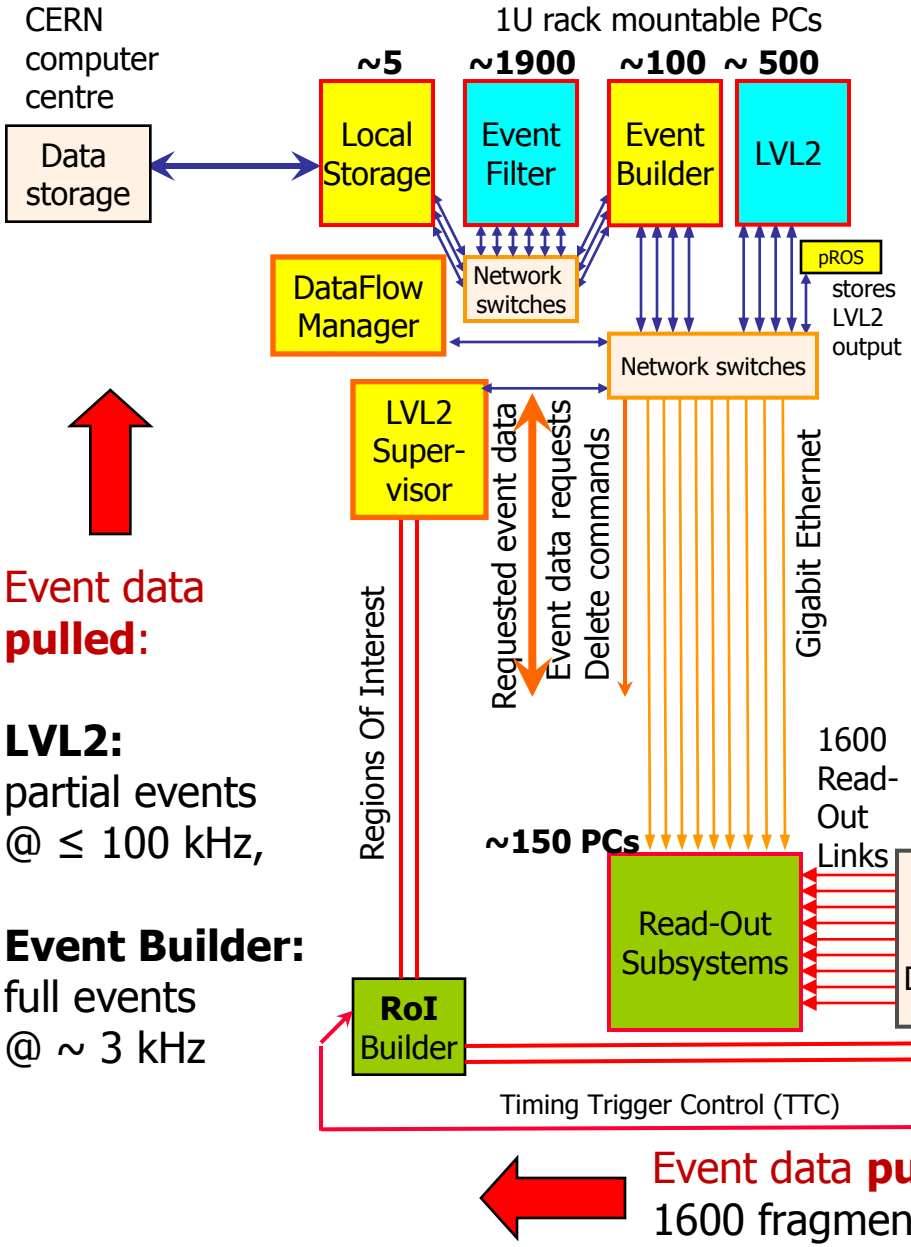
Three trigger levels to reduce the initial bunch crossing rate to a rate acceptable for data taking

- **LVL1** hardware trigger
- **LVL2** PC farm
  500 1U PCs – multi core
  reconstruction of data within **Regions of Interest** as defined by LVL1
- **Event Filter** PC farm
  1900 1U PCs – multi core
  reconstruction of the **whole event**

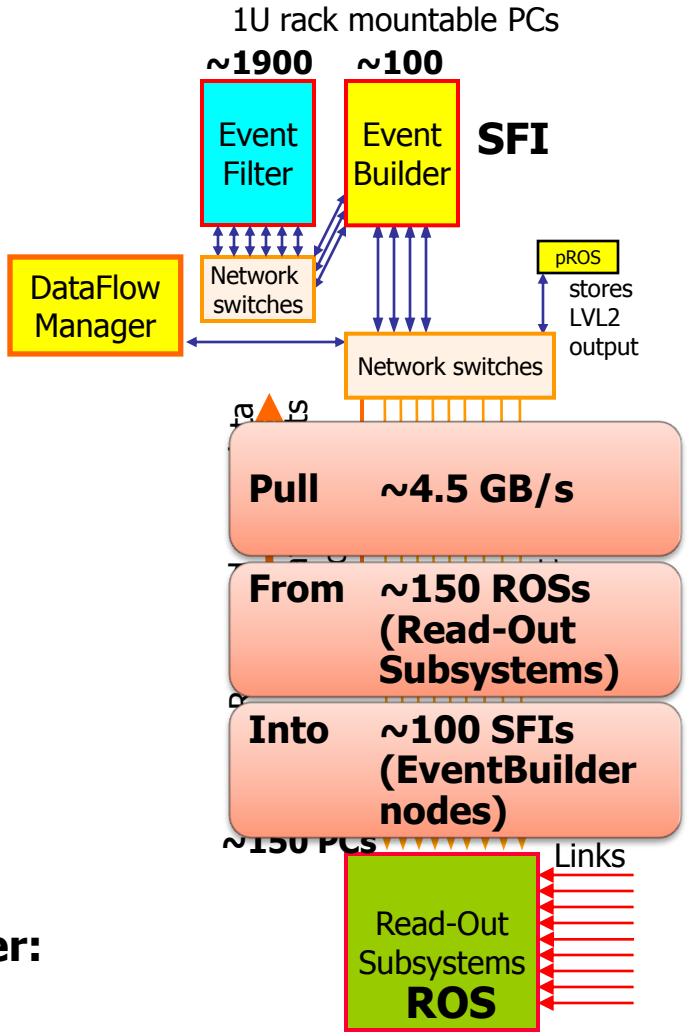


See Imma Riu talk
On ATLAS TDAQ Integration
TDAQ-Sys01

ATLAS Trigger & Data Acquisition

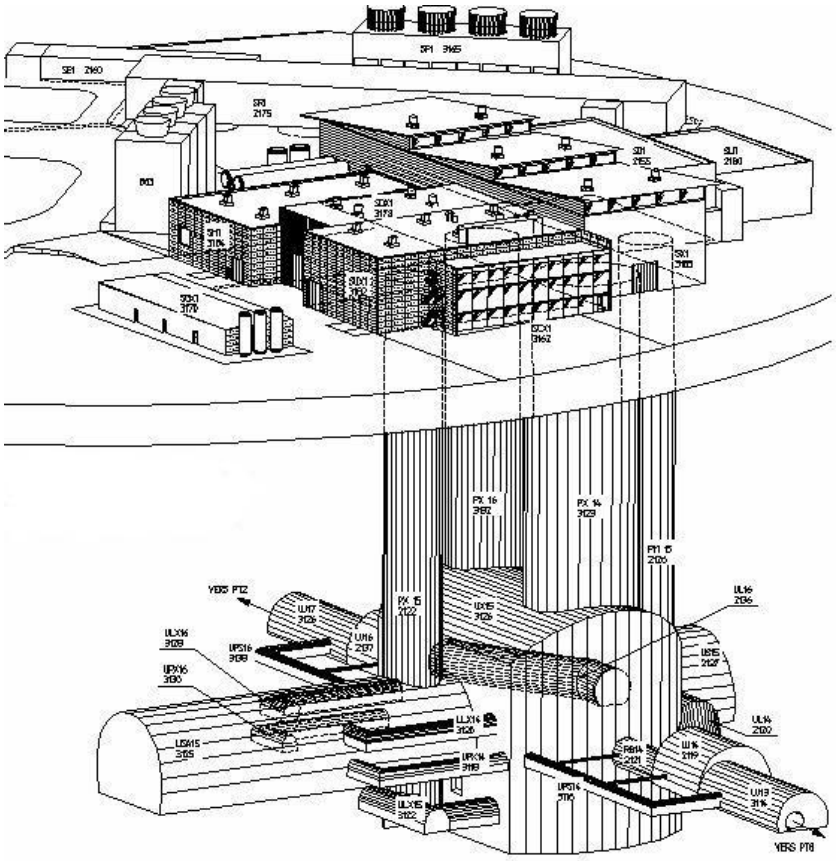


ATLAS Event Builder



Event data **pulled**:

Event Builder:
full events
@ ~ 3 kHz

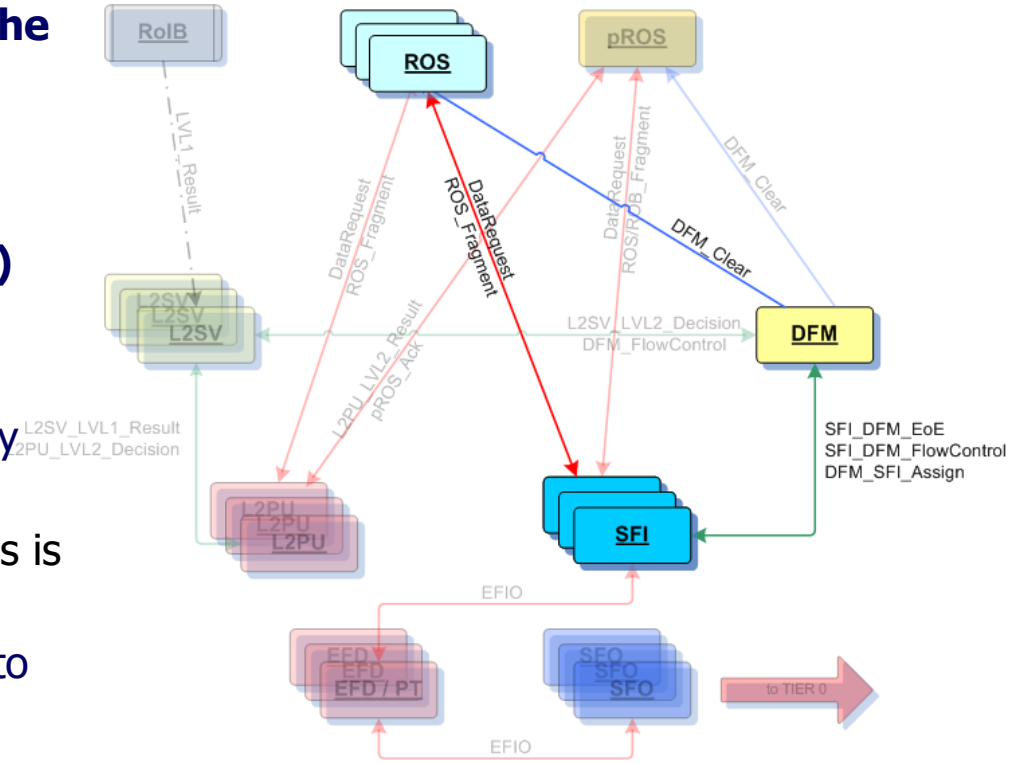


~1.5 MB event size

Event data **pushed** @ ≤ 100 kHz,
1600 fragments of ~ 1 kByte each

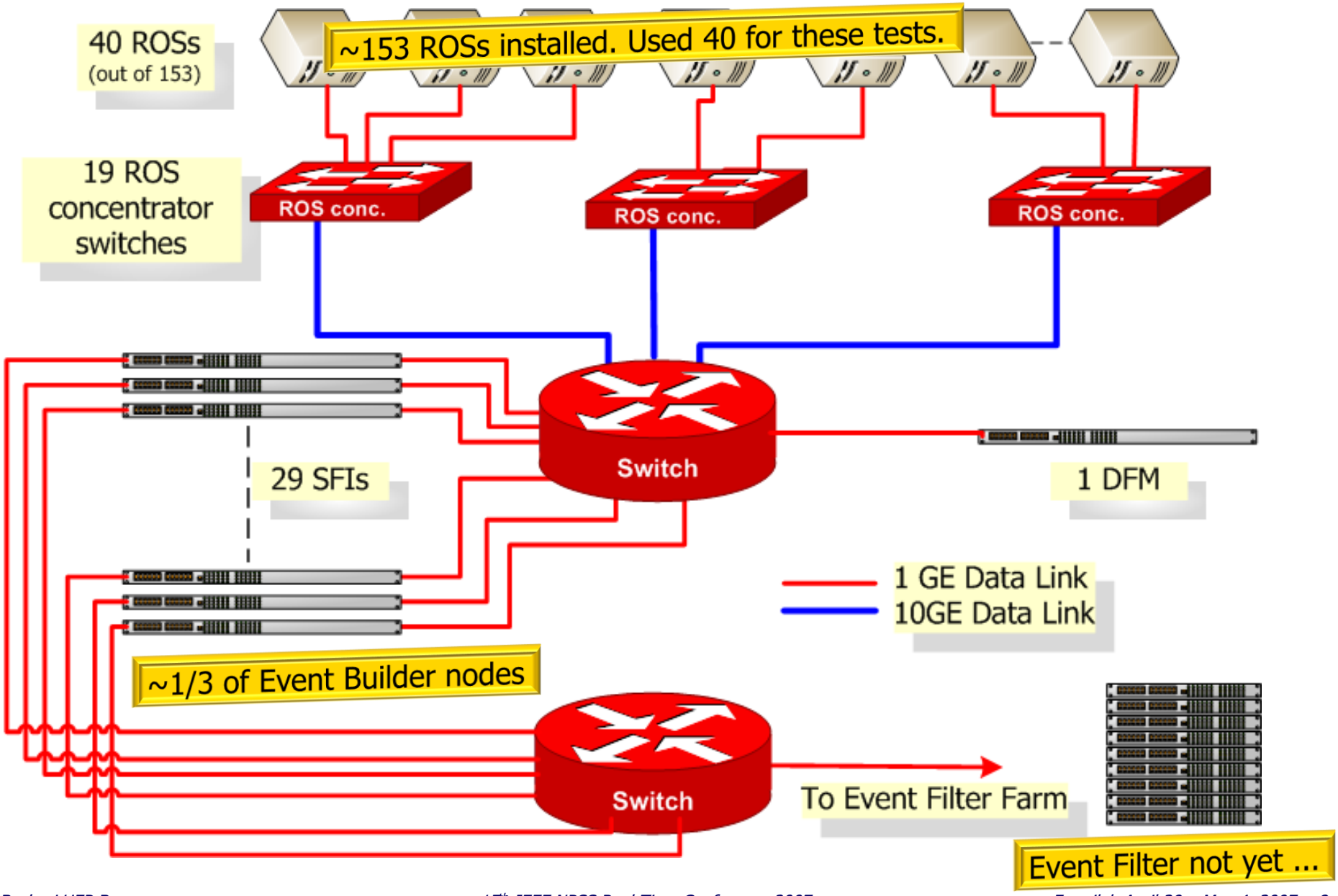
The Event Builder Pull protocol

- **The DFM recevies a trigger via the network**
  - From LVL2 (usually)
  - From LVL1 (commissioning)
  - **Self-triggering (these tests)**
- Upon a trigger, the DFM assigns one free SFI to build the event
- The SFI sends data requests to every ROS
  - Number of outstanding requests is limited → traffic shaping
- The ROS send their ROS_Fragment to the requesting SFI → and keep the data
- The SFI receives the ROS_Fragement
  - Or re-asks for the fragment again if transfer failed (timeout)
- The SFI builds the event from all ROSs
- The SFI informs the DFM that the event is finished
- The DFM sends a clear message to all ROSs



- **Network Protocols used**
  - UDP / IP for data requests and data replies
  - UDP / IP multicast for the DFM clear messages
  - TCP / IP for data flow commands
  - Possibility to use TCP / IP everywhere

Eventbuilder Topology in Spring 2007:

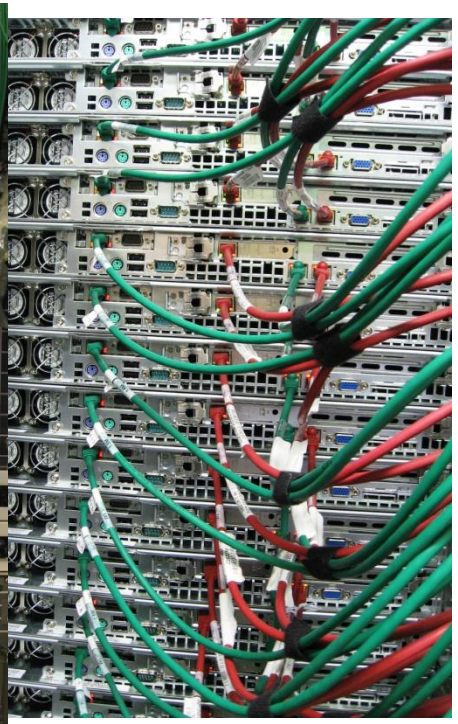


Read-Out subsystem

- **153 ROS PCs installed**
  - 40 used for these tests
- 4U, 19" rack mountable PC
- Motherboard: Supermicro X6DHE-XB
- CPU: One 3.4 GHz Xeon
- Hyper threading not used
- uni-processor kernel
- RAM: 512 MB
- Network:
  - **2 GB onboard**
    1 used for control network
  - **4 GB on PCI-Express card**
    1 used for LVL2 data
    1 used for event building
- Redundant power supply
- Network booted (no local hard disk)
- Remote management via IPMI

The Event Builder Node: SFI

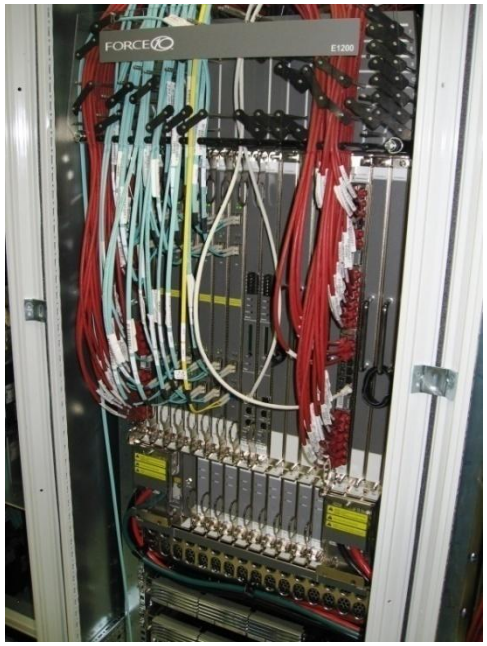
- **32 SFI PCs installed**
  - Final system ~100 SFIs
  - 29 SFIs used in these tests
- 1U, 19" rack mountable PC
- Motherboard: Supermicro H8DSR-i
- CPU: AMD Opteron 252 2.6 GHz
- SMP kernel
- RAM: 2 GB
- Network:
  - **2 GB onboard**
    1 used for control network
    1 used for data-in
  - **1 GB on PCI-Express card used for data-out**
  - 1 dedicated IPMI port
- Cold-swappable power supply
- Network booted
- Local hard disk to store event data; only used for commissioning
- Remote management via IPMI



The DataFlow Manager: DFM

- **12 DFM PCs installed**
  - **Final system needs 1 DFM**
  - 12 DFMs
    - run up to 12 TDAQ partitions in parallel
    - useful during commissioning
- **Same PC as for SFI**
- Network:
  - **2 GB onboard**
    1 used for control network
    1 used for data network
    1 dedicated IPMI port
- Cold-swappable power supply
- Network booted
- Local hard disk (not used)
- Remote management via IPMI

The Switches



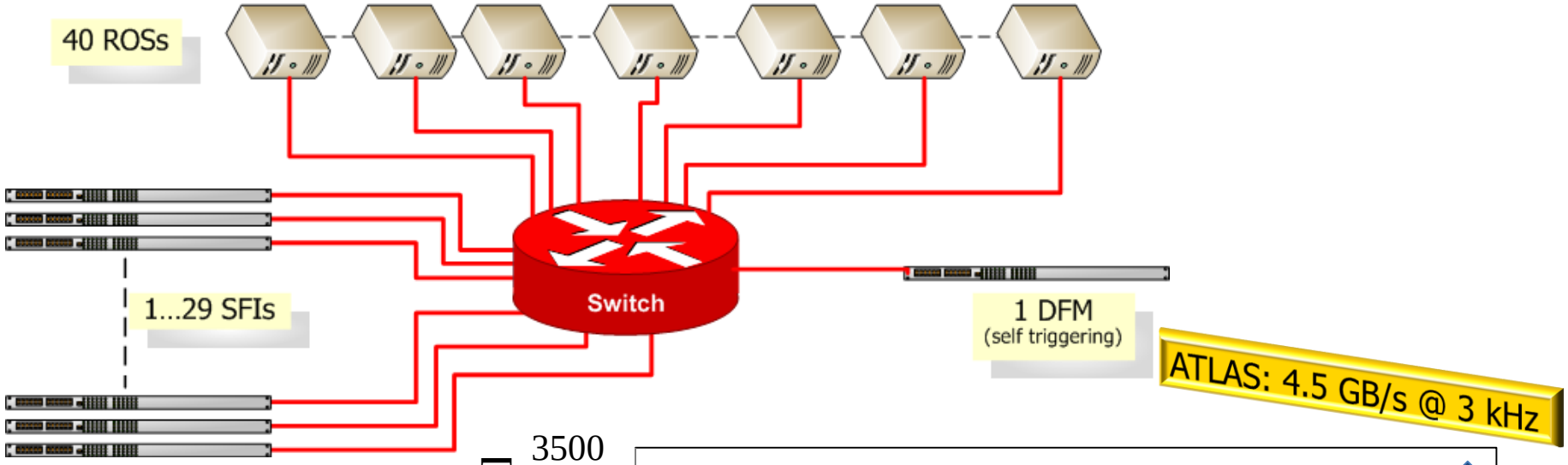
- **Force10 E1200**
  - 6 blades x 4 optical 10GE ports
  - 2 blades x48 copper GE ports
  - Up to 14 blades
    1260 GE ports total
    672 GE ports @ line speed
- **Data network**
  - Event builder traffic
  - LVL2 traffic

- **Force10 E600**
  - Up to 7 blades
    630 GE ports total
    336 GE ports @ line speed
- **Data network**
  - To Event Filter

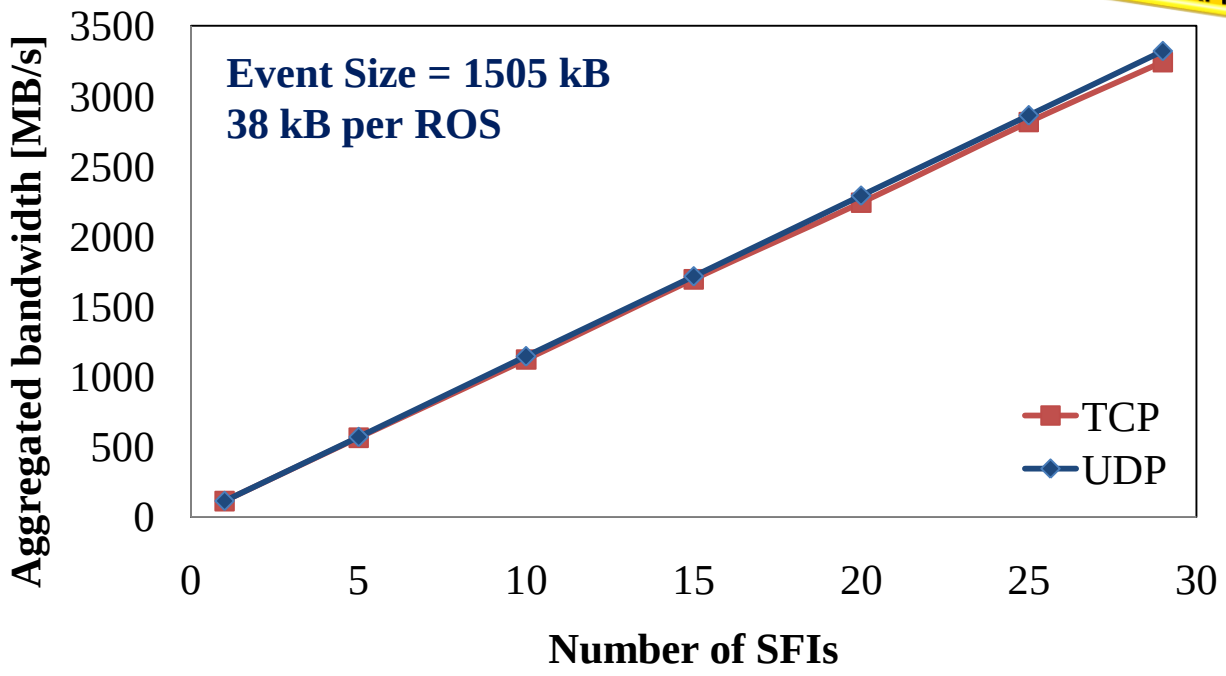
See Silvia Batraneanu's talk On ATLAS TDAQ Networks TD-DAQ02

- **Force10 E600**
  - Up to7 blades
    630 GE ports total
    336 GE ports @ line speed
- **Control network**
  - Run Control
  - Databases
  - Monitoring samplers

Measuring the scaling properties

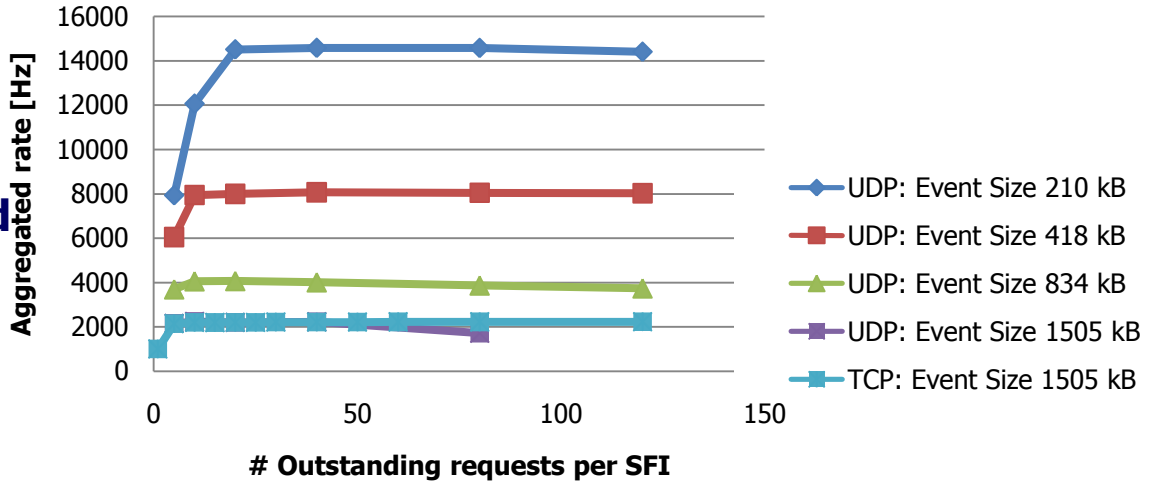


- **Perfect scaling observed up to 29 SFIs @ Event size of 1.5 MB**
  - 3.3 GB/s aggregated bandwidth
  - 2.2 kHz EB rate
  - Single SFIs close to GE speed
    → 114 MB/s per SFI
    → 78 Hz per SFI



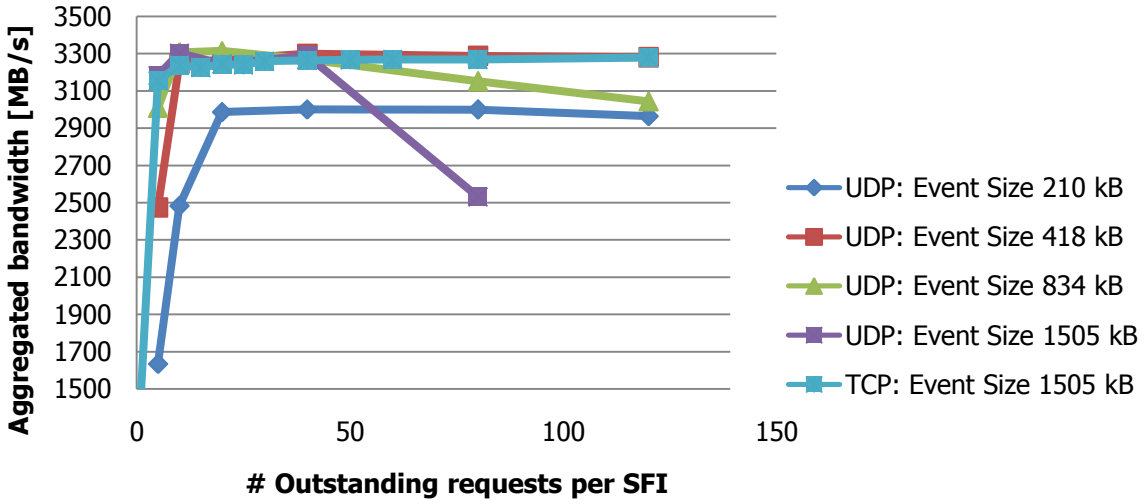
Traffic Shaping

- **Traffic shaping is achieved by limiting the number of outstanding requests per SFI**

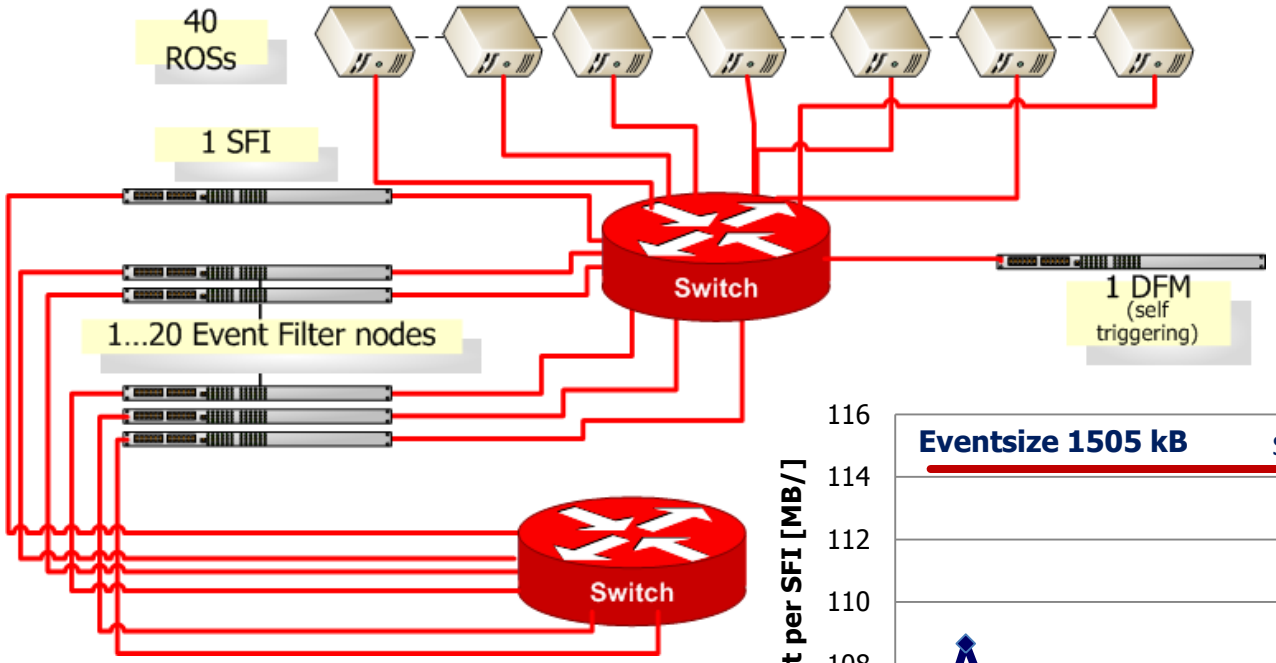


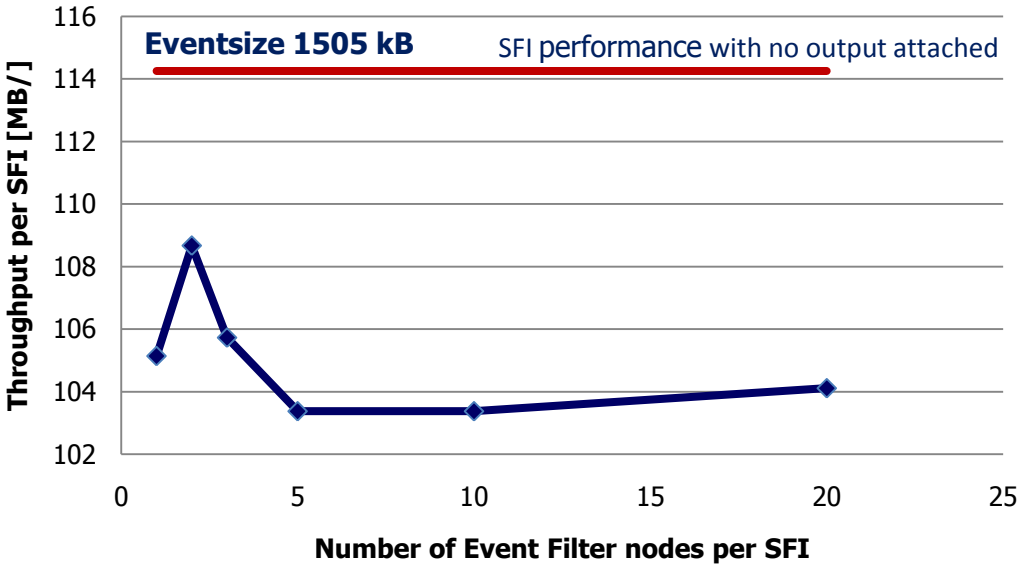
- For big event sizes and large number of outstanding requests, the aggregated bandwidth drops

→ packet loss and subsequent re-ask of data fragment



Building Events and sending them to Event Filter

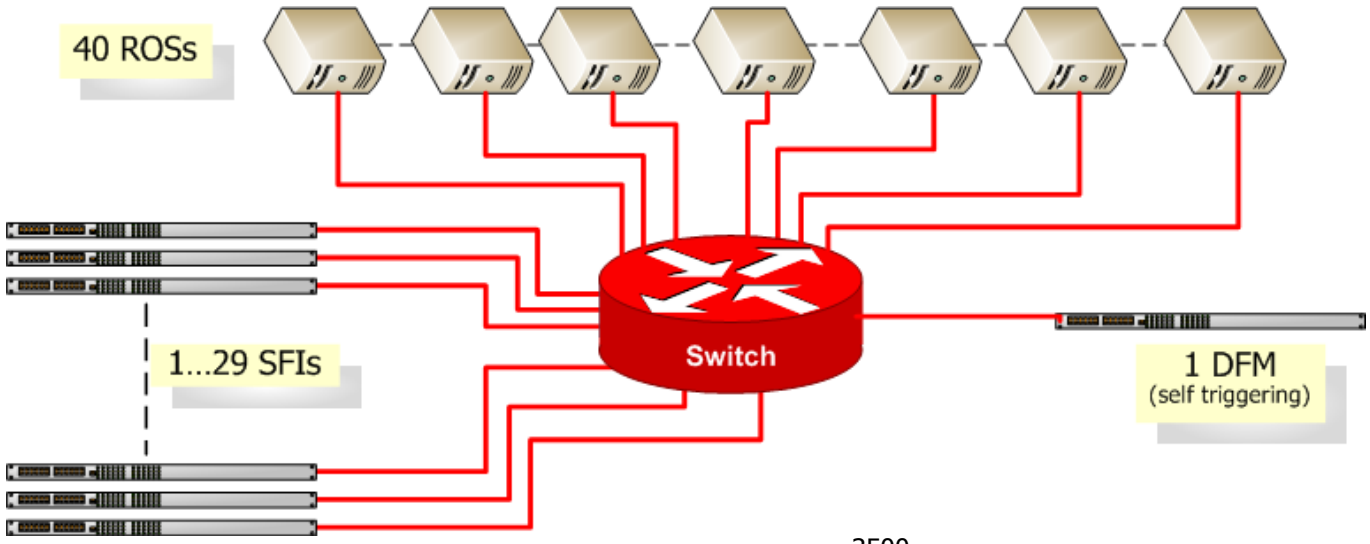




- **SFI PCs can also act as Event Filter nodes**
- Pulling data out of SFI as fast as possible
  - Zero CPU time spent for event processing

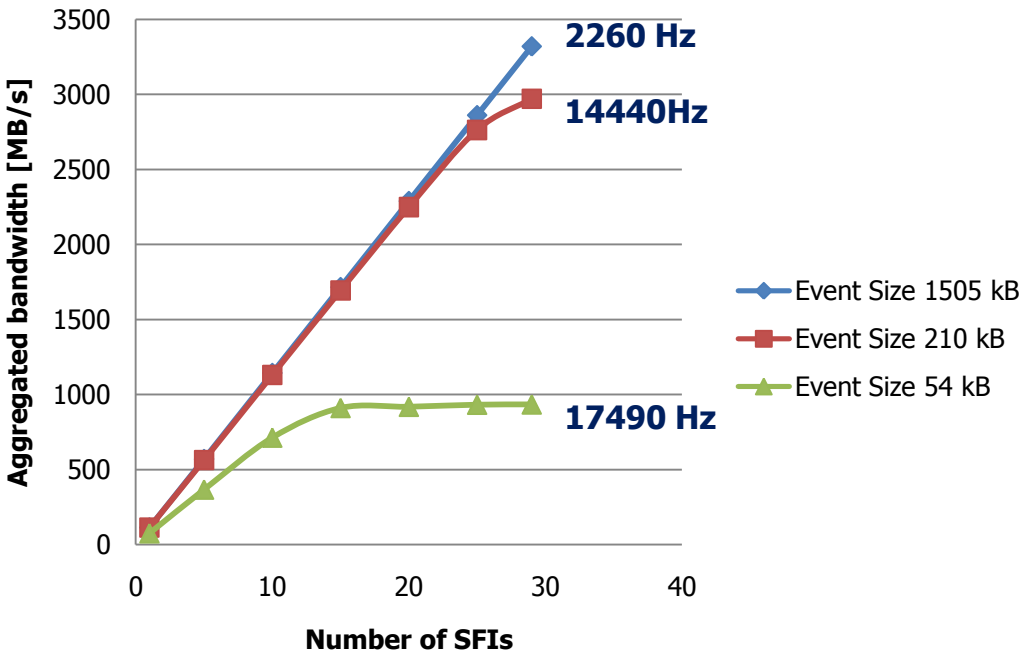
~10% degradation when sending data to Event Filter

Reaching the limit of the Read-Out subsystem



- Setting the **event size to very small values** is equivalent to a high event building frequency
  - The ROS becomes rate limited
  - Adding more SFIs does not increase the aggregated bandwidth any further
- **No problem for building events of 1.5 MB @ 3 kHz**

Event building rate:

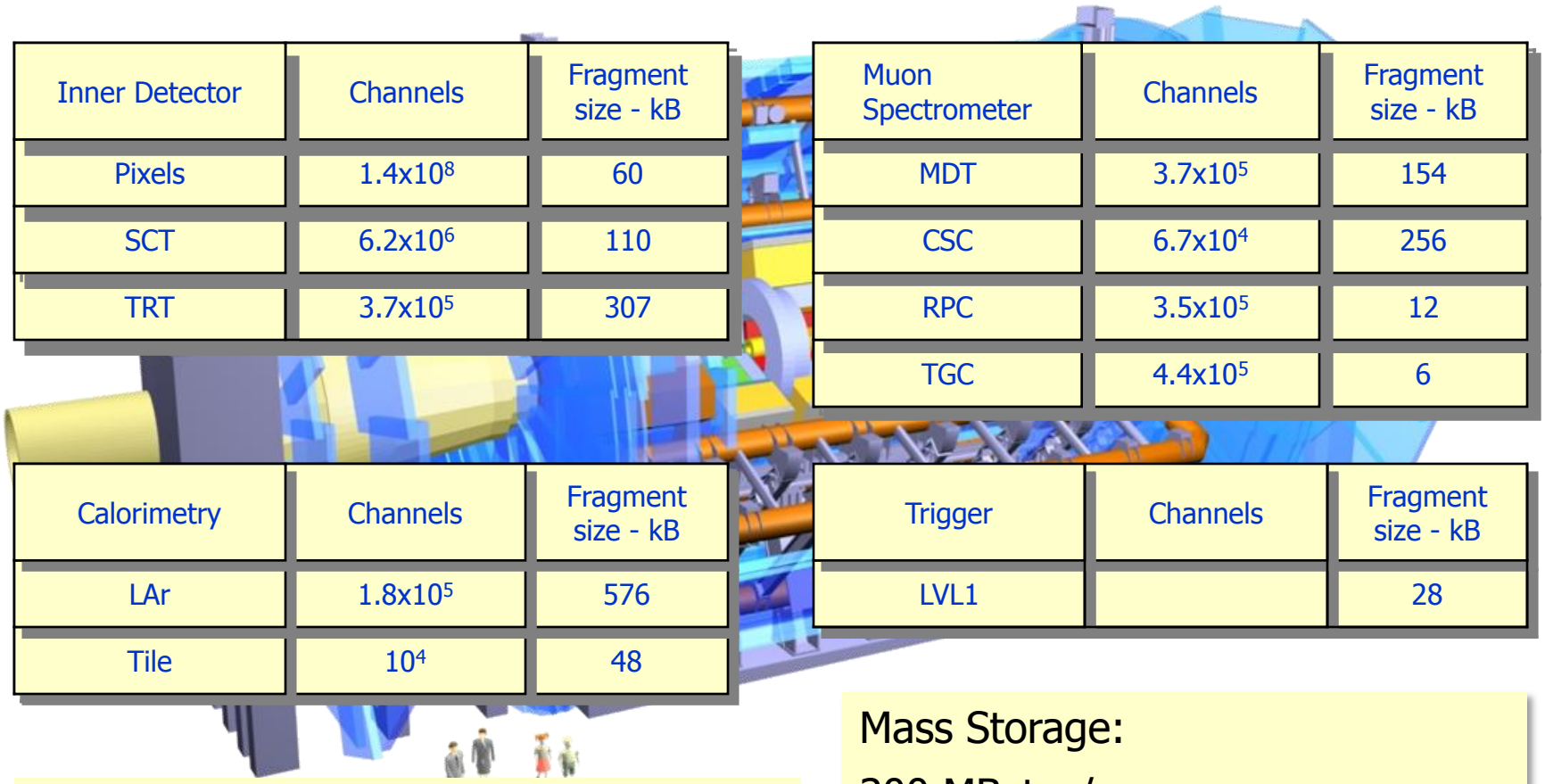


Conclusions

- **1/3 of the ATLAS Event Builder is installed**
  - All 153 Read-Out subsystems (ROSs) installed
    - All 149 ROSs are used for detector commissioning plus 4 spares
    - 40 ROSs used for these tests
  - 32 Event Builder nodes (SFIs) installed
    - 29 SFIs used for these tests
- **The ATLAS Event Builder is based on a pull protocol**
  - Data Flow Manager (DFM)
    - receives triggers from LVL2, LVL1 or self-triggering
    - Load-balances the SFI farm
  - Event Builder node (SFI)
    - Requests Data fragments from ROSs
    - In case of packet loss, data fragments can be re-asked
  - **Will use UDP / IP for requesting data and for sending data**
    - can also use TCP / IP
- **Have reached 2/3 of required bandwidth and rate with 1/3 of event builder nodes**
  - **29 SFIs can do 2.2 kHz @ 1.5 MB per event → 3.3 GB/s**
  - Expect 10% degradation when data is also sent to Event Filter
  - But → no LVL2 traffic added yet….
- **It looks very promising to go even beyond ATLAS requirements — if needed**

BackUp

Atlas Event Size



Inner Detector	Channels	Fragment size - kB
Pixels	1.4×10^8	60
SCT	6.2×10^6	110
TRT	3.7×10^5	307

Muon Spectrometer	Channels	Fragment size - kB
MDT	3.7×10^5	154
CSC	6.7×10^4	256
RPC	3.5×10^5	12
TGC	4.4×10^5	6

Calorimetry	Channels	Fragment size - kB
LAr	1.8×10^5	576
Tile	10^4	48

Trigger	Channels	Fragment size - kB
LVL1		28

Atlas event size: 1.5 Mbytes
140 Mio Channels
organized into **~1600 Readout Links**

Mass Storage:

300 MBytes/sec

☛ 3 PetaBytes/year
for offline analysis