The ATLAS Event Builder

W.Vandelli*
CERN – Physics Department/ATD

On behalf of ATLAS TDAQ DataFlow

* *This research project has been supported by a Marie Curie Early Stage Research Training Fellowship of the European Community's Sixth Framework Programme under contract number (MRTN-CT-2006-035606)*

ATLAS TDAQ DataFlow

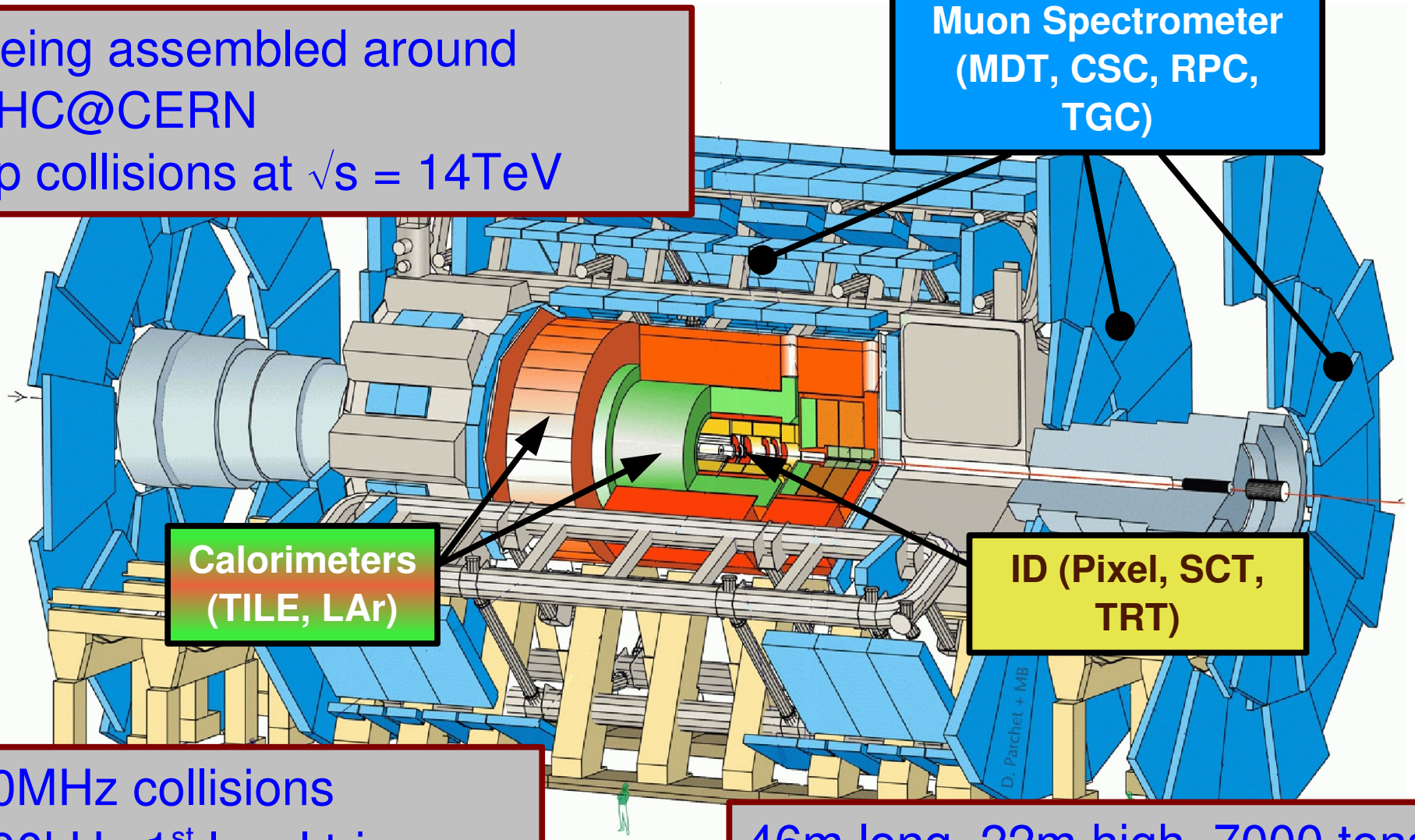
H.P. Beck[1], M. Abolins[2], A. Battaglia[1], R. Blair[3], A. Bogaerts[4], M. Bosman[5], M. Ciobotaru[6], R. Cranfield[7], G. Crone[8], J. Dawson[3], R. Dobinson[4†], M. Dobson[4], A. Dos Anjos[9], G. Drake[3], Y. Ermoline[2], R. Ferrari[10], M.L. Ferrer[11], D. Francis[4], S. Gadomski[1], S. Gameiro[4], 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

ATLAS Experiment

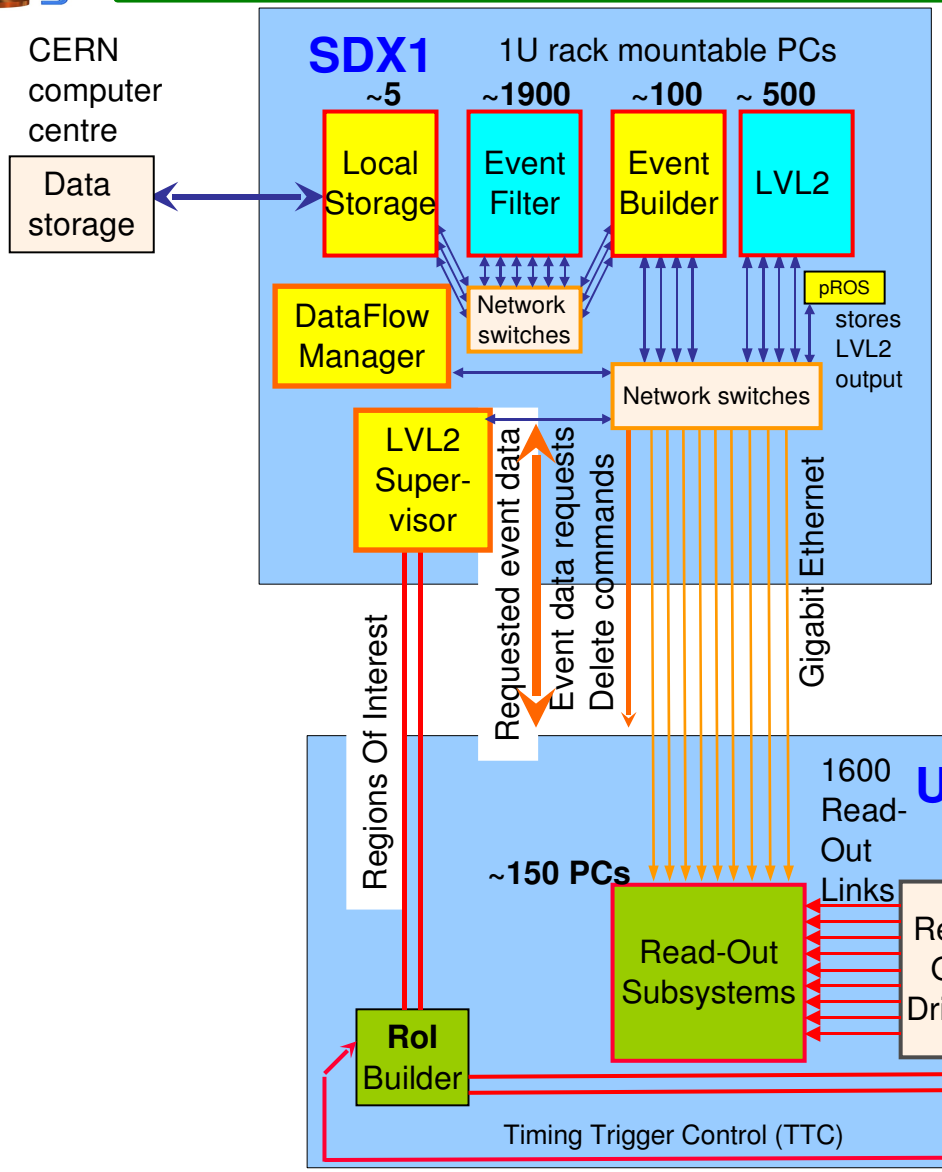
Being assembled around LHC@CERN
pp collisions at $\sqrt{s}$ = 14TeV



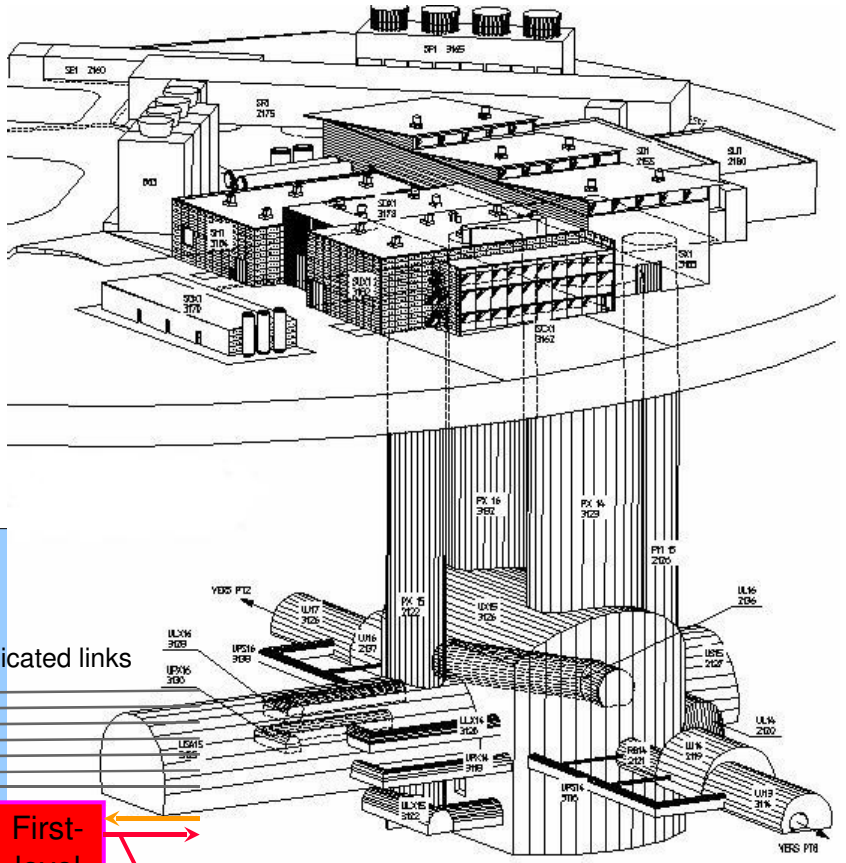
40MHz collisions
100kHz 1st level trigger
O(100Hz) stored events

46m long, 22m high, 7000 tons
140M channels

ATLAS TDAQ

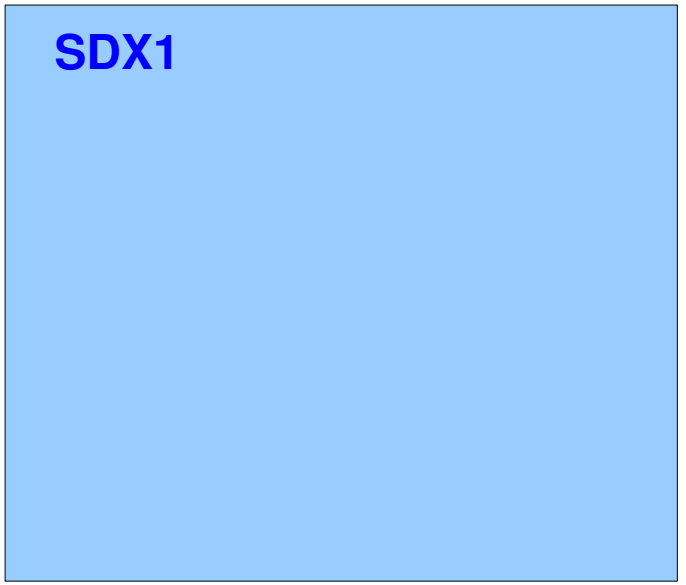


Mostly C++ applications running on Linux

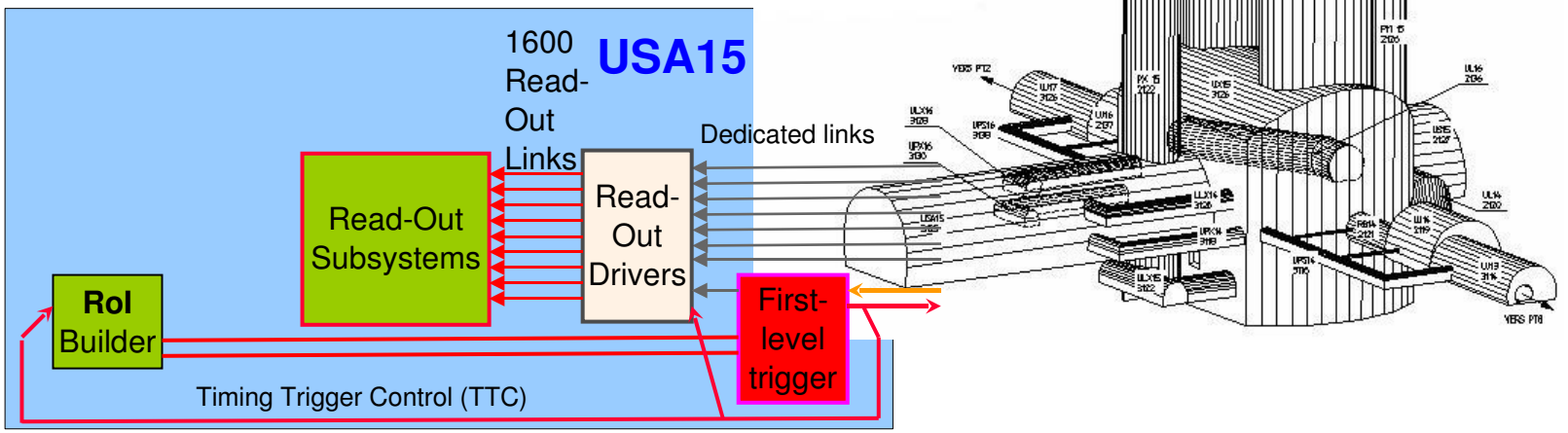


Event data **pushed** @ ≤ 100 kHz,
1600 fragments of ~ 1 kByte each

Read-Out System

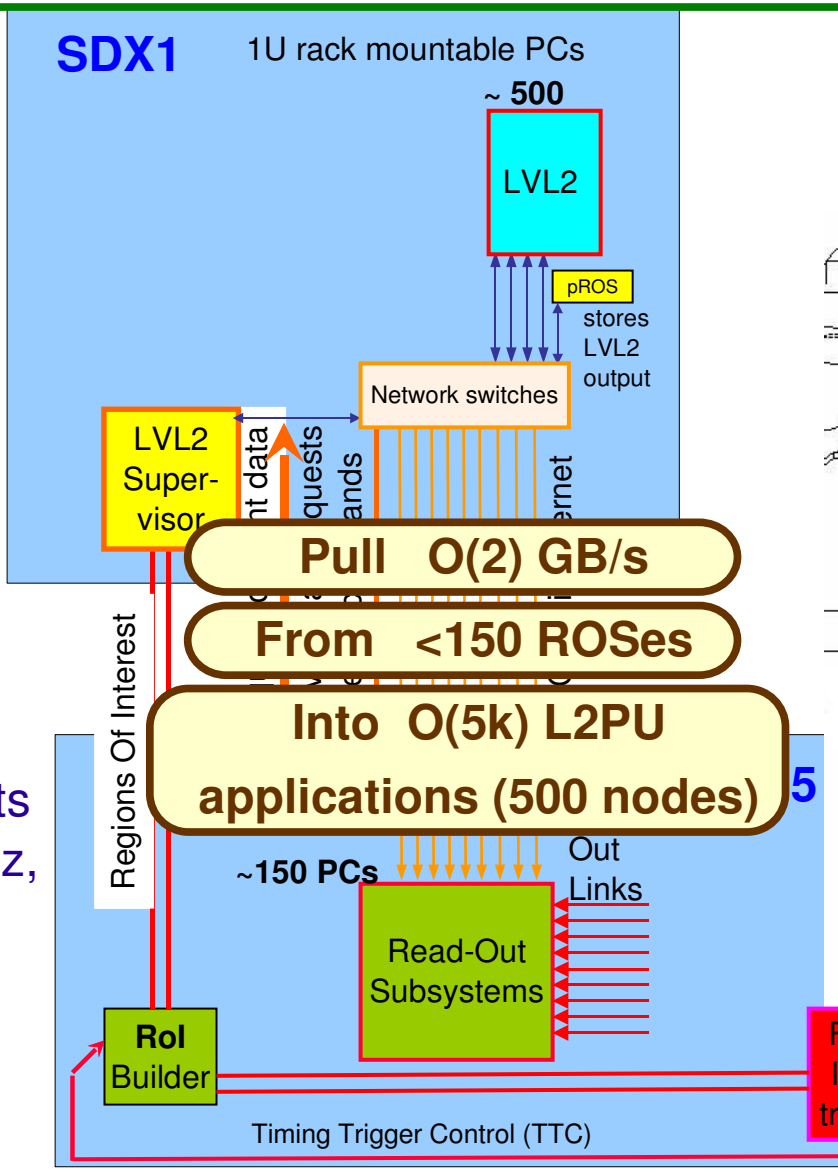


- 150 PCs housing custom PCI boards
  - 1600 optical readout links (ROL)
- ROSes hold the data till the LVL2 decision
  - serve data to LVL2 and Event Builder

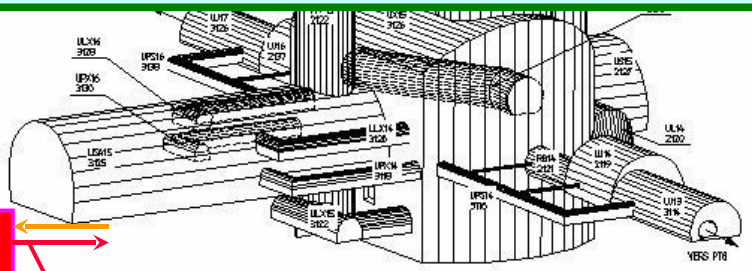


Event data **pushed** @ ≤ 100 kHz,
1600 fragments of ~ 1 kByte each

LVL2 System



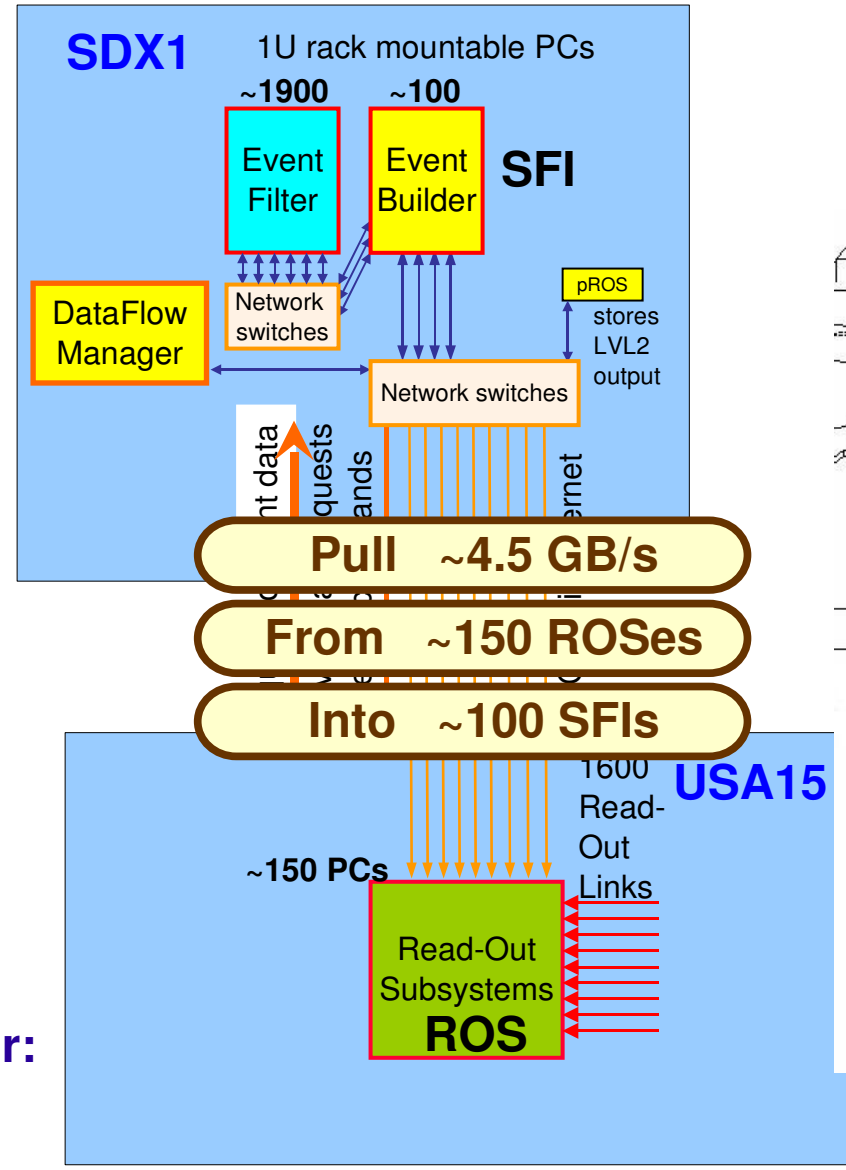
- **Reconstruction in limited detector regions (RoI) defined by the LVL1**
- **Highly non-uniform ROS request rate**
  - **"Hot ROSes" ~12kHz**
- **#RoI and #ROL/RoI**
  - **depend on physics, LVL1, algorithms**
  - **long-tailed distributions**
- **Expected O(2) ROL/ROS/RoI**

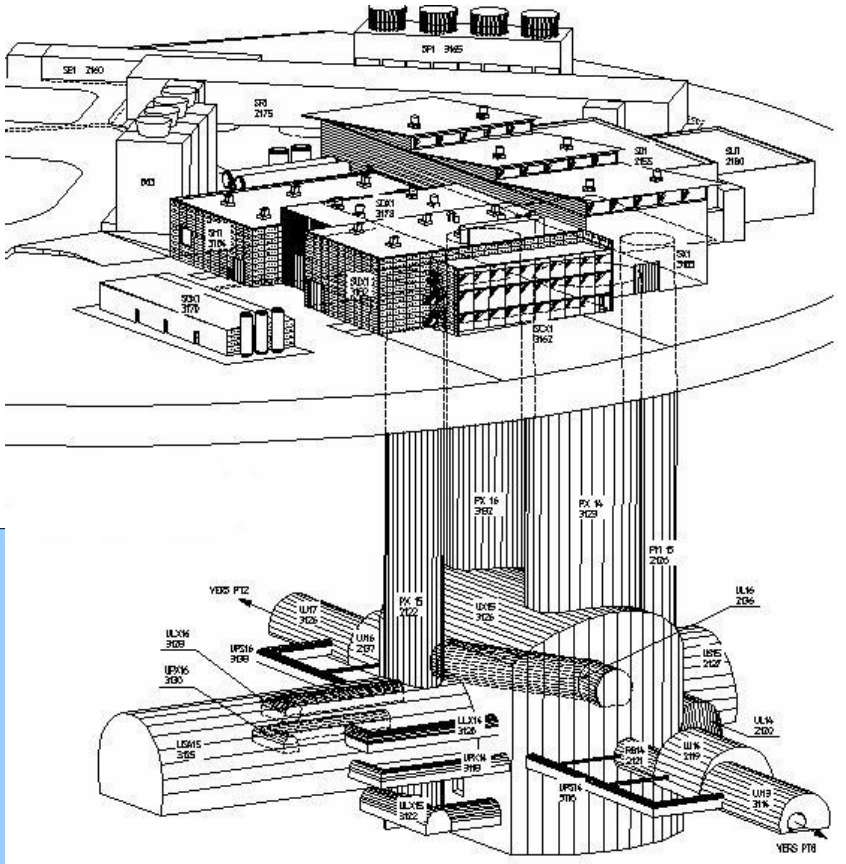


LVL2:
partial events
@ ≤ 100 kHz,

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

ATLAS Event Builder Commitment





Event data **pulled**:

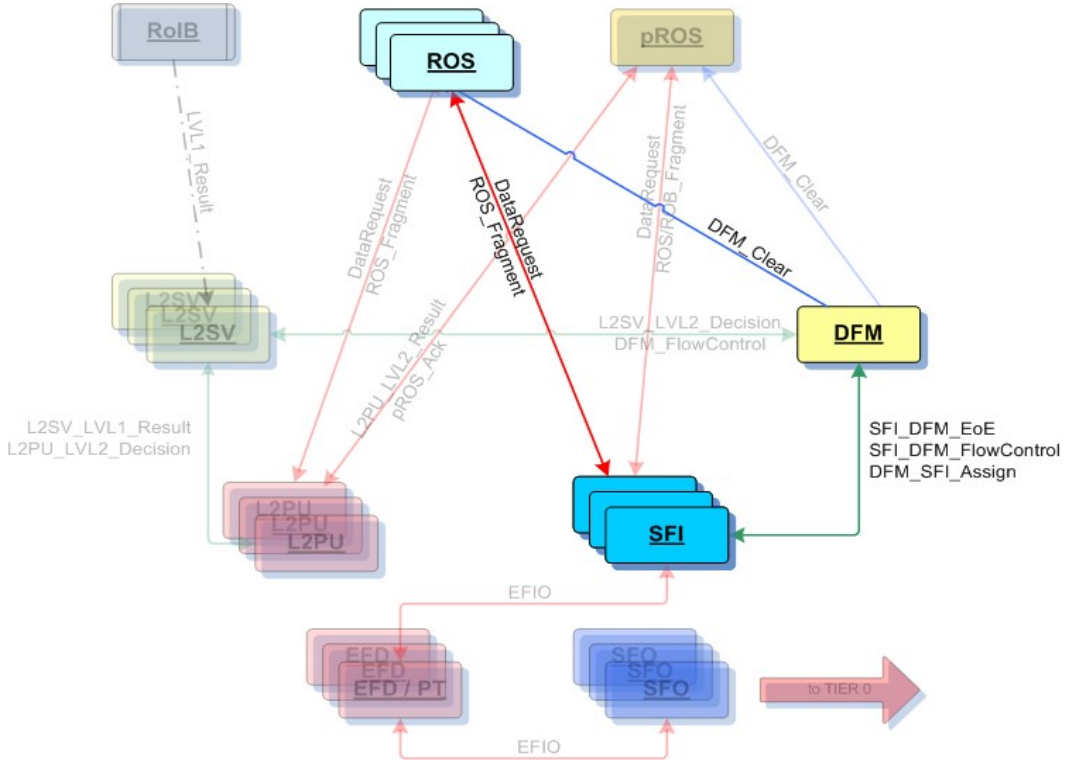
Event Builder:
full events
@ ~ 3 kHz

~1.5 MB event size

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

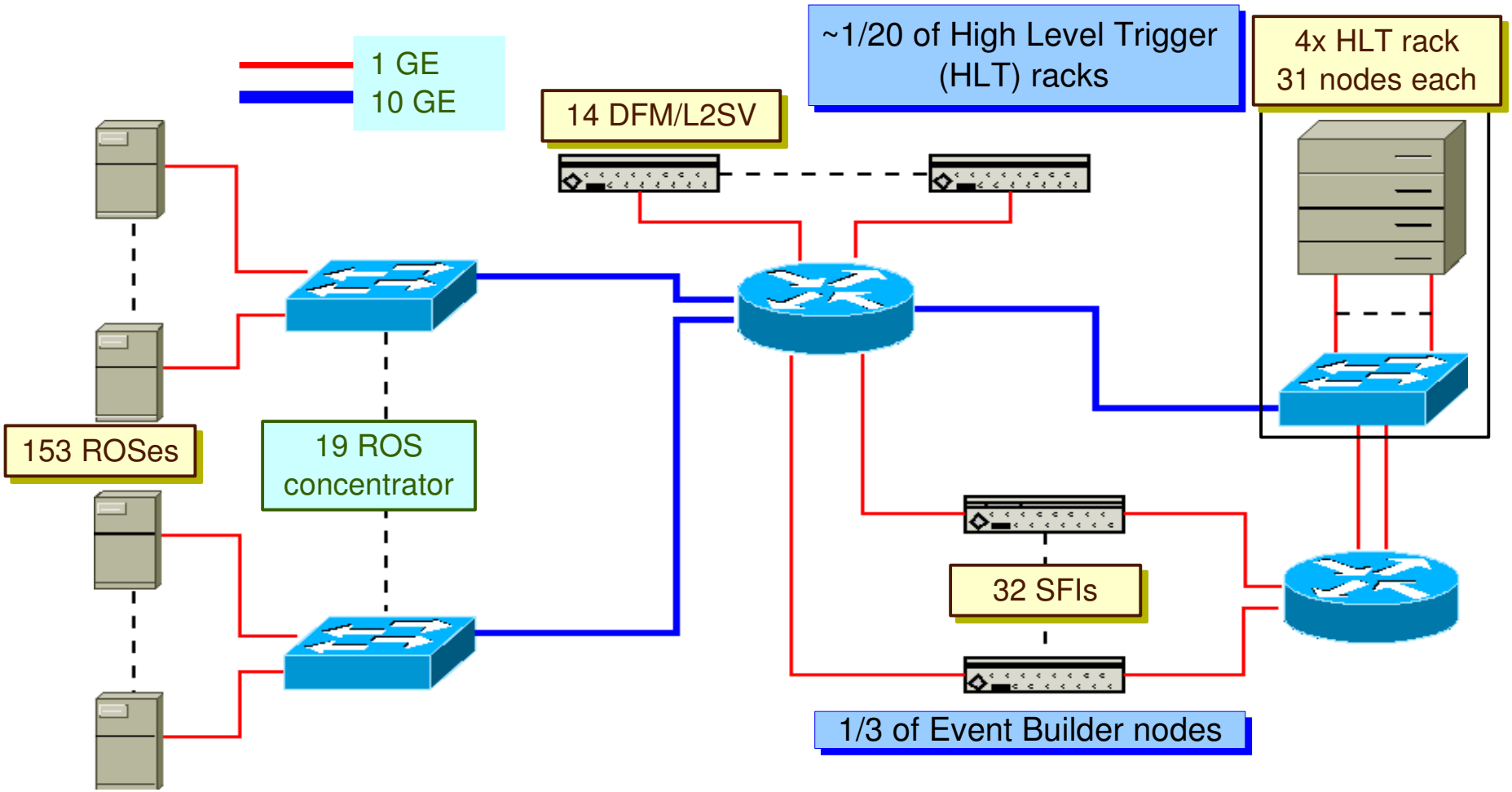
Event Builder Protocol

- DFM is the "orchestra leader" of the EB system
  - receives trigger via the network
  - assigns events to the SFIs, load balancing the farm
  - handles End-Of-Event
  - sends clear messages to the ROSes
- SFI is the event builder application
  - asks all the ROSes for data fragments
  - builds a complete event
  - serves the Event Filter
- SFI features
  - Traffic shaping: limited number of outstanding requests
  - Re-asks missing fragments
  - Monitoring server



- **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

Presently Installed Hardware

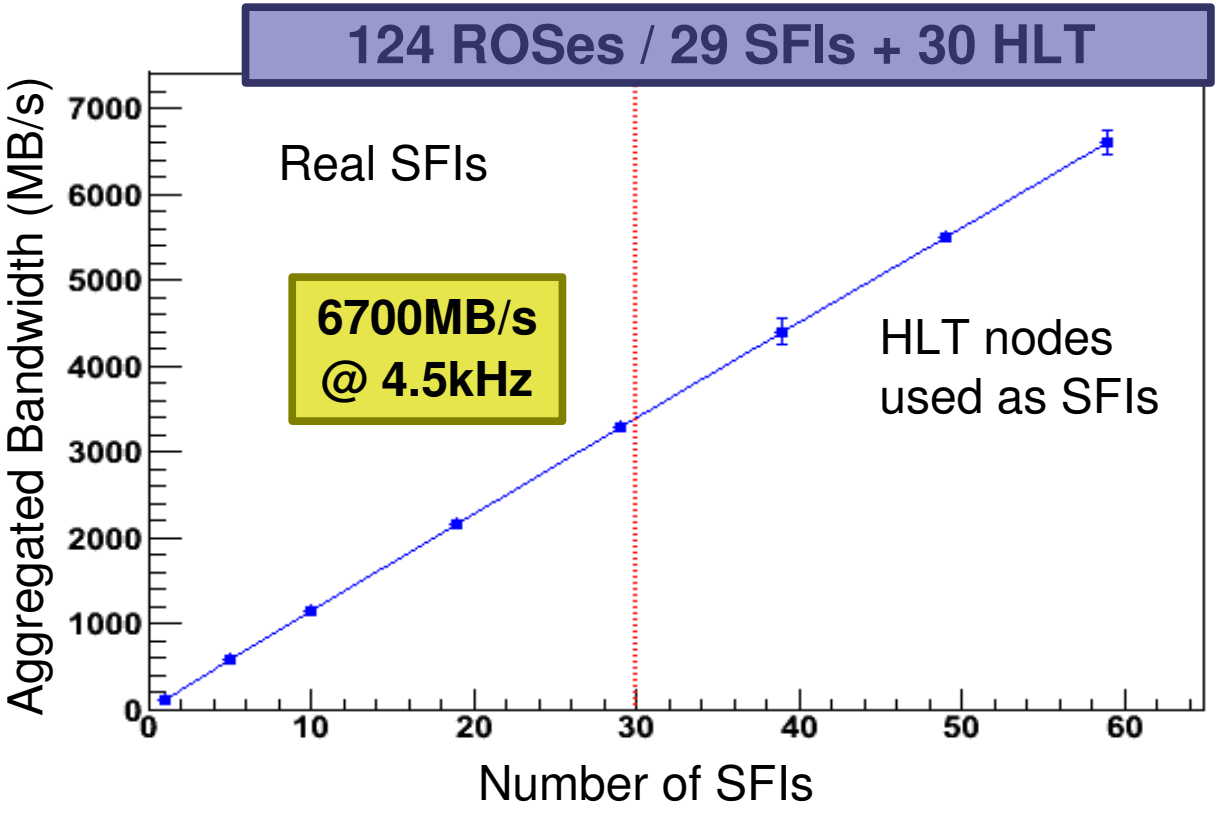


- SFI/DFM: dual AMD Opteron 252 2.6 GHz
- HLT: dual Intel Xeon E5320 quad-core 1.86GHz

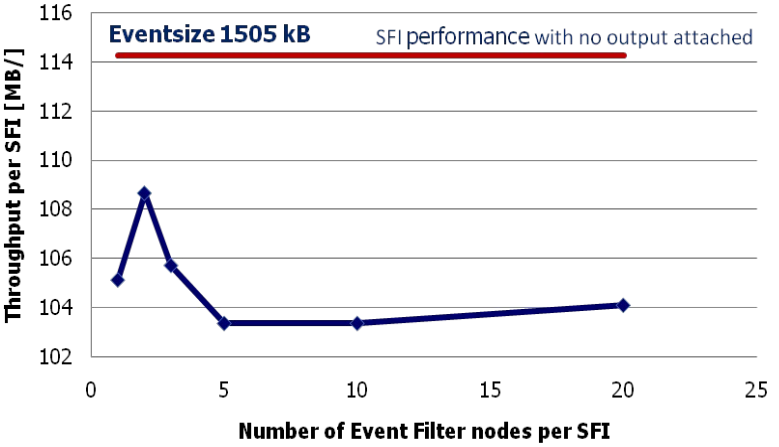
Central & local file server, online service and monitoring nodes not shown

Event Builder Scaling Properties

- Extend test capabilities running SFI application on HLT nodes
- Perfect scaling up to 59 SFIs
  - 1.5 MB event size
  - 6.5GB/s @ 4.5kHz
- SFI application can roughly exploit the full GE link
  - 114MB/s @ 76 Hz (SFI)
  - 112MB/s @ 75 Hz (HLT)
- Multi-core HLT nodes have good SFI performance
  - Double SFI approach
- Very promising result, BUT

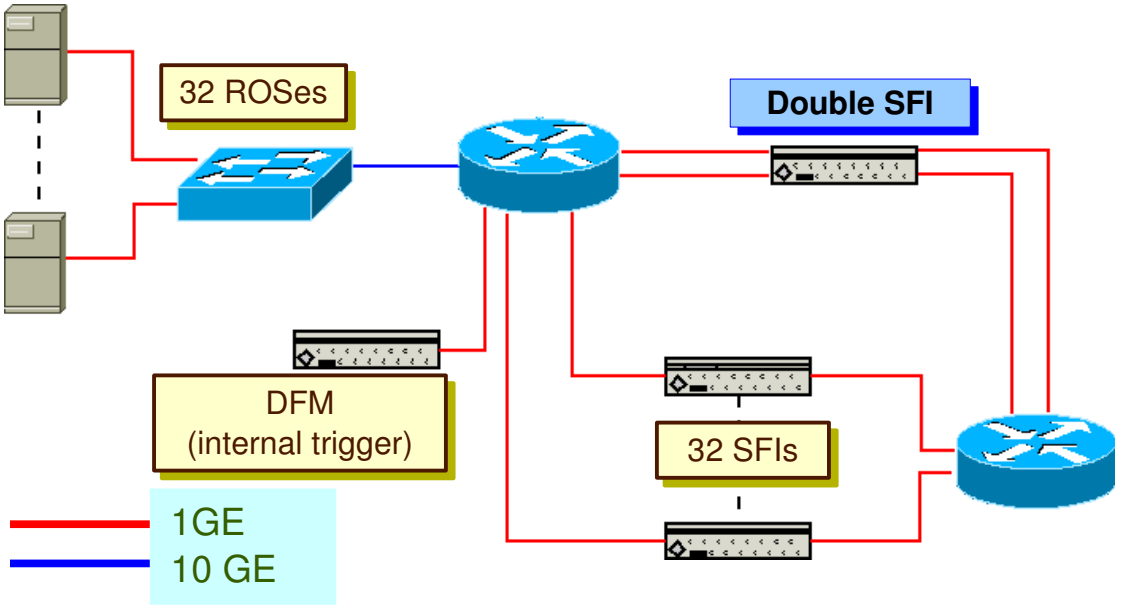


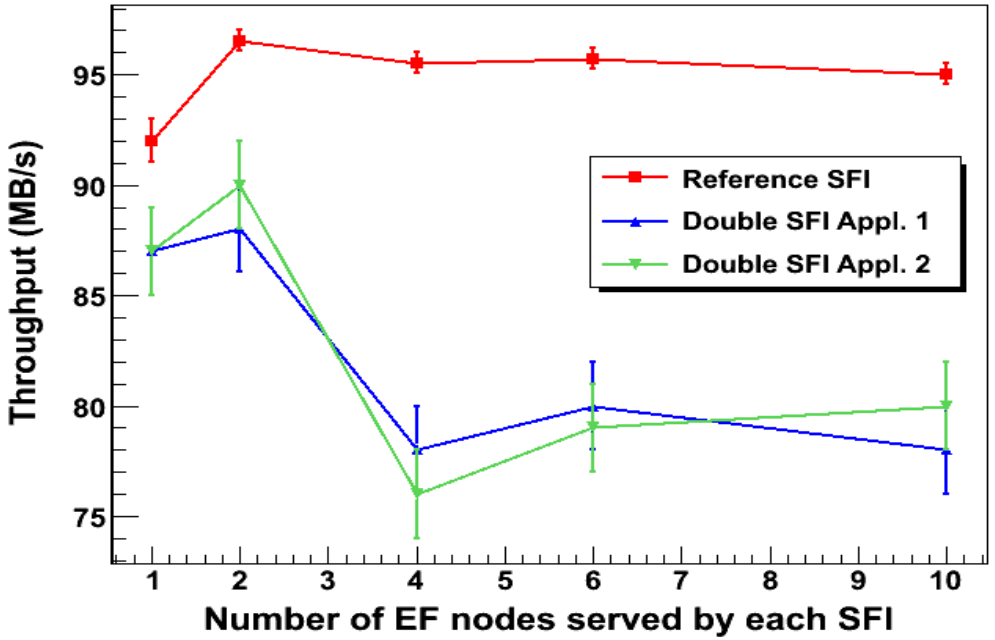
10% performance decrease when sending data to Event Filter



No LVL2 load on system (ROSes & network) during the test

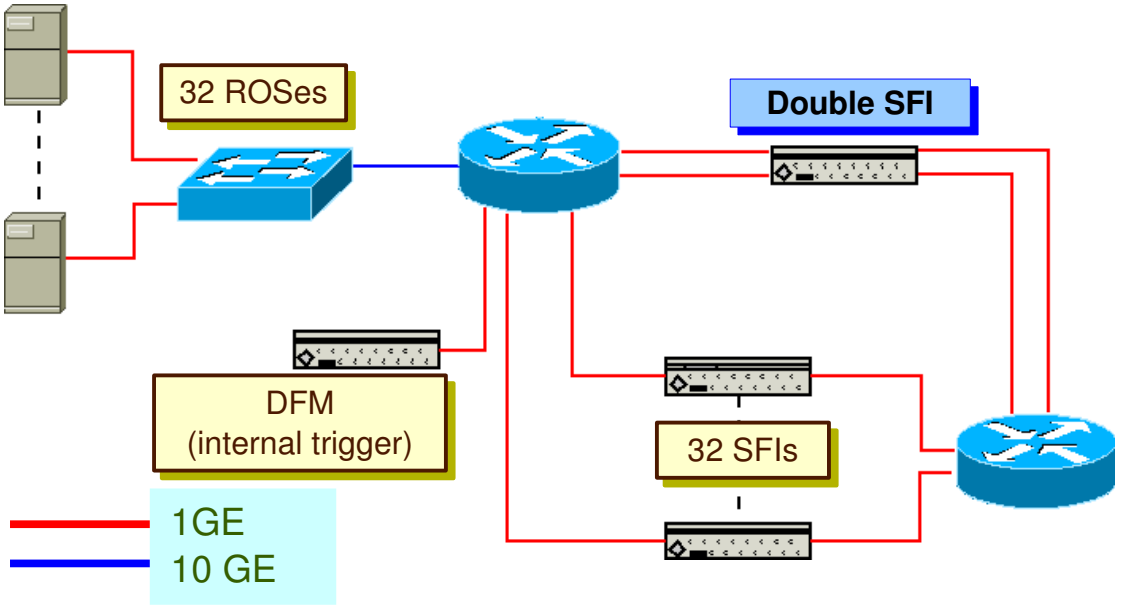
Double SFI Approach

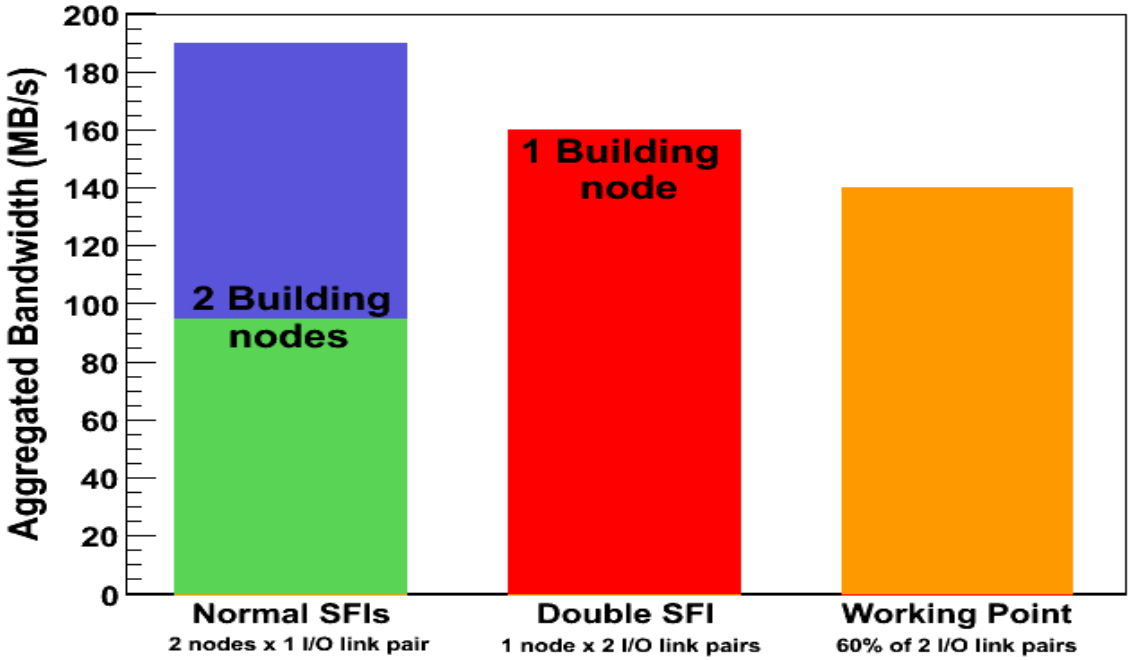




- Exploit the availability of multi-core processors on the same node running multiple SFI applications
  - extended network capability needed
- HLT node with quad-NIC board
  - dual-CPU quad-core 1.86GHz
  - 2 bonded interfaces toward ROSes
  - 2 bonded interfaces toward the EF
- Throughput with 32 ROSes, 1.1 MB event
  - Reference SFI 95 MB/s
  - Double SFI 2 x 80 MB/s = 160 MB/s

Double SFI Approach

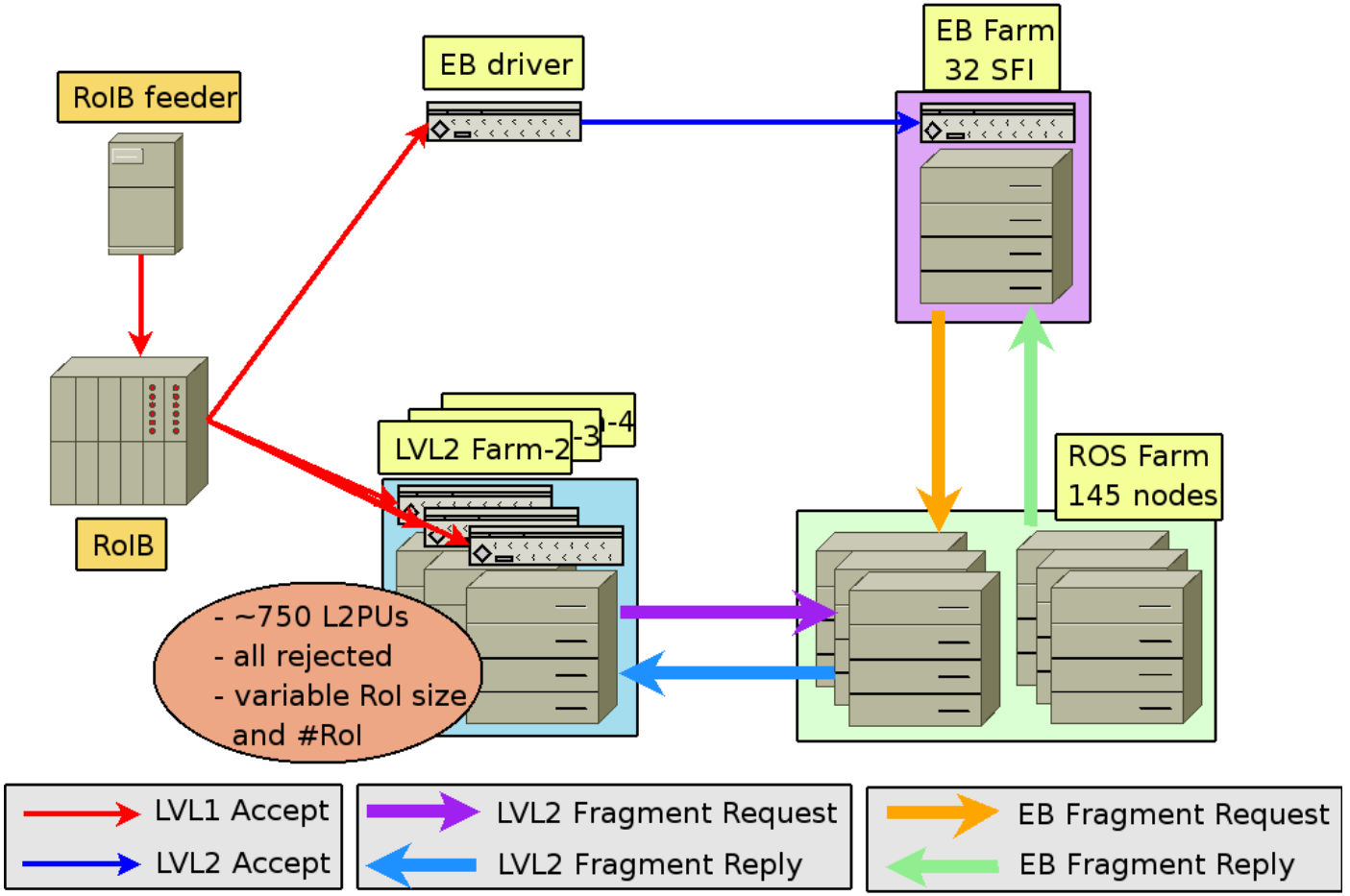




- Exploit the availability of multi-core processors on the same node running multiple SFI applications
  - extended network capability needed
- HLT node with quad-NIC board
  - dual-CPU quad-core 1.86GHz
  - 2 bonded interfaces toward ROSes
  - 2 bonded interfaces toward the EF
- Throughput with 32 ROSes, 1.1 MB event
  - Reference SFI 95 MB/s
  - Double SFI 2 x 80 MB/s = 160 MB/s
- Good performance, limited by the (output) bonding efficiency
  - Input only: 2 x 114 MB/s = 228 MB/s
  - Foreseen working point <70 MB/s per SFI
  - Far from the CPU limit: dual-CPU dual-core

Including LVL2 load in the system

- Due to the limited computing resources cannot use real LVL2 algorithms
- Simulate the LVL2 load on network and ROSes with LVL2 processing unit (L2PU) running dummy algorithm.
  - RoI size
  - #RoIs
- Custom topology with EB driven by an independent, non-requesting LVL2 farm
  - decouple EB and LVL2
  - independently tune corresponding working points



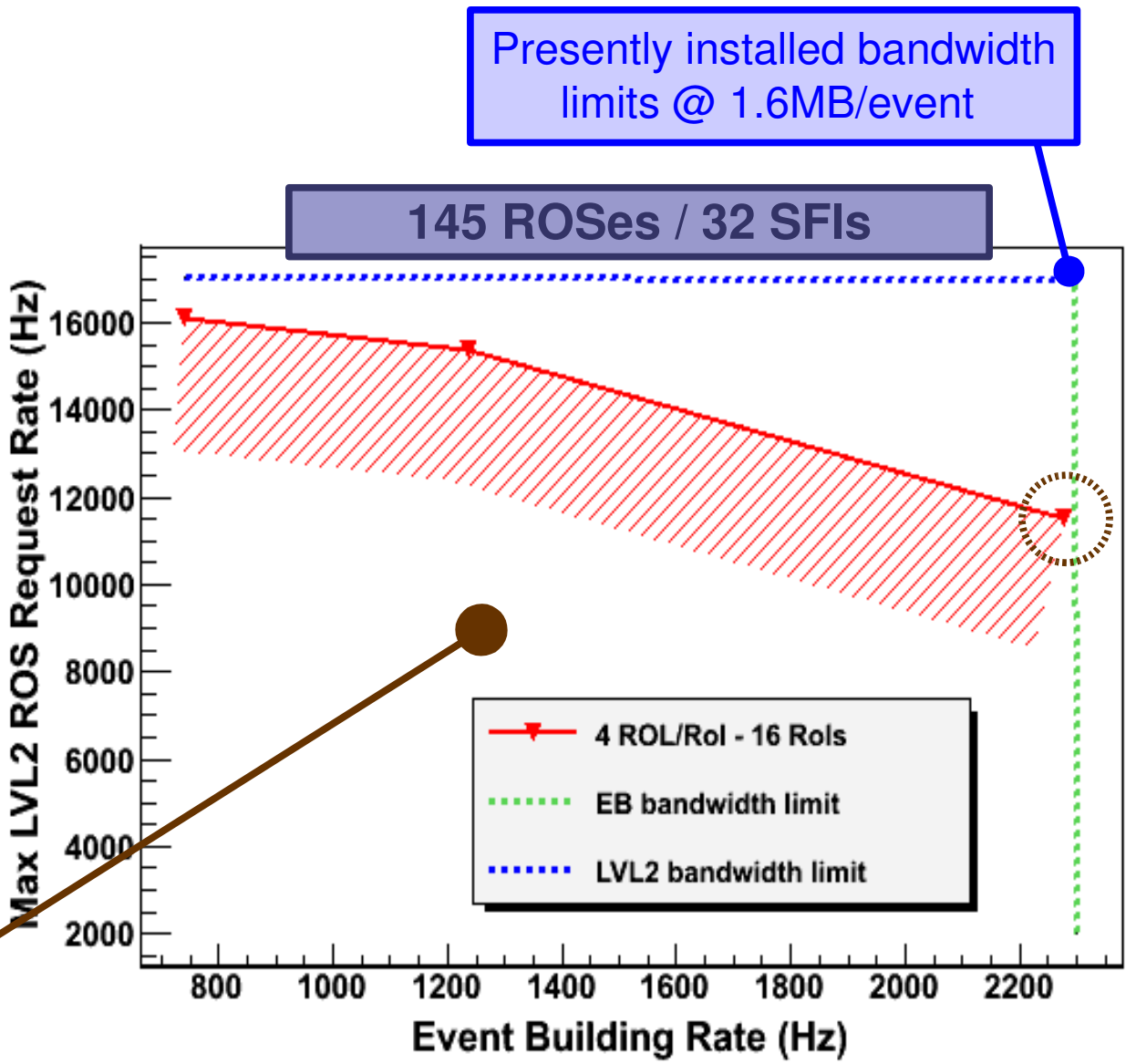
Try to reach the highest L2 request rate on the ROSes, at a given RoI size, varying the number of RoIs per event

Measure the effects on the EB and on the ROSes

LVL2 and Event Builder

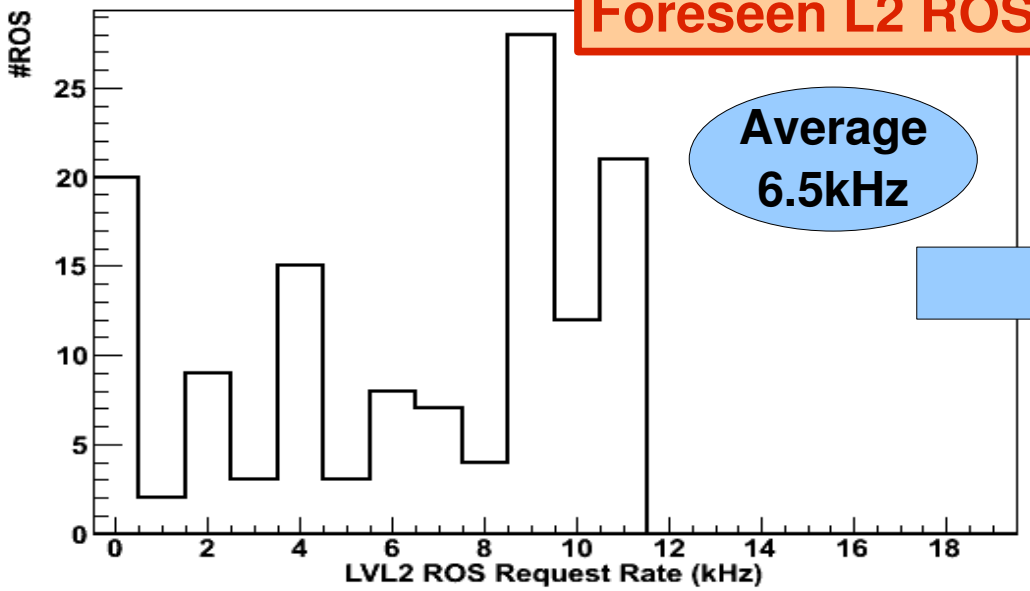
- Not able to completely decouple Event Building and LVL2
- Observed decrease in the LVL2 rate is due to a decrease of LVL1 rate
- System limits not exposed
  - Limited by the presently installed hardware
  - Correlation effect in the RoIB system

The system can always sustain the Event Builder in the explored space



Closer look to the ROSes

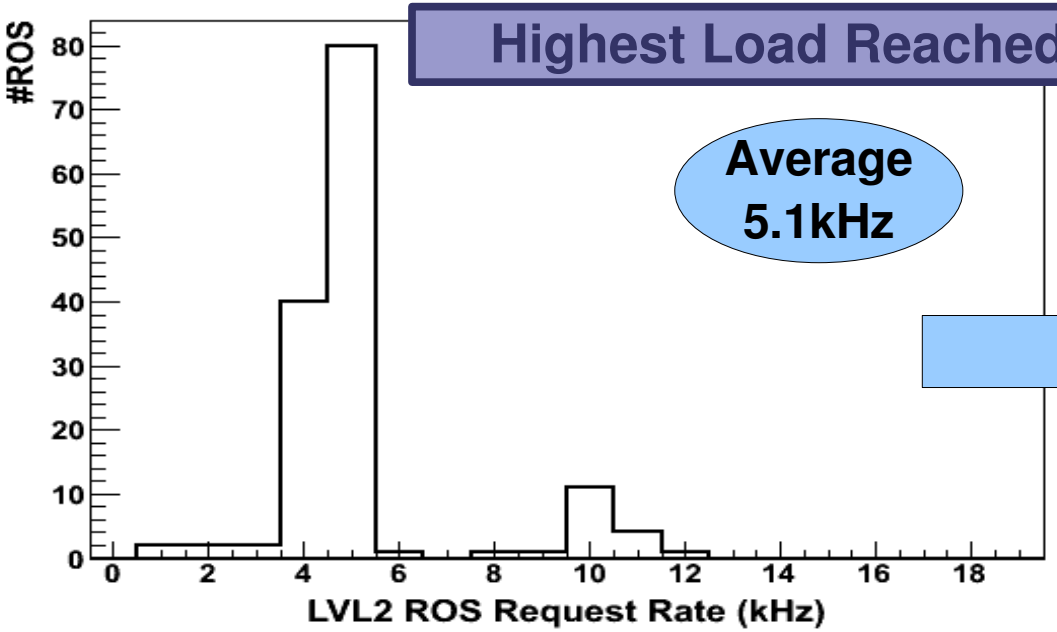
Foreseen L2 ROS Request Rate



Average 6.5kHz

LVL2 Traffic O(2) GB/s
EB Traffic 4.5 GB/s

Highest Load Reached

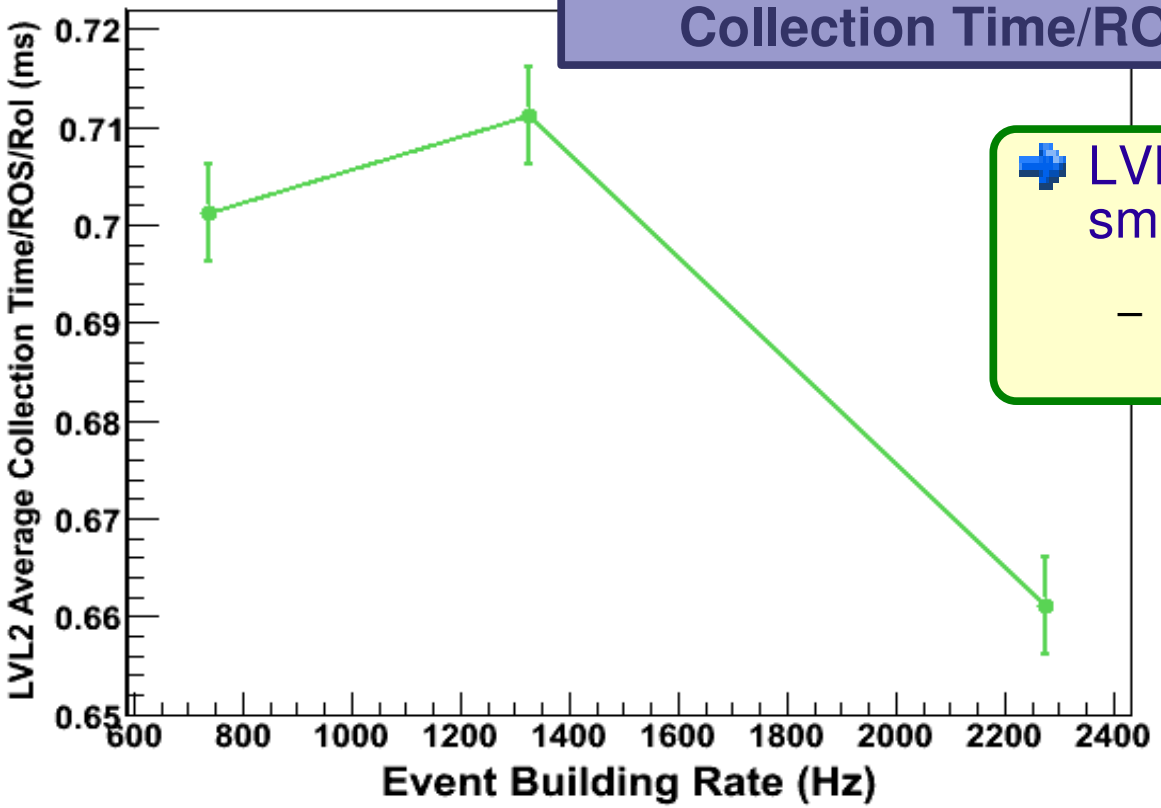


Average 5.1kHz

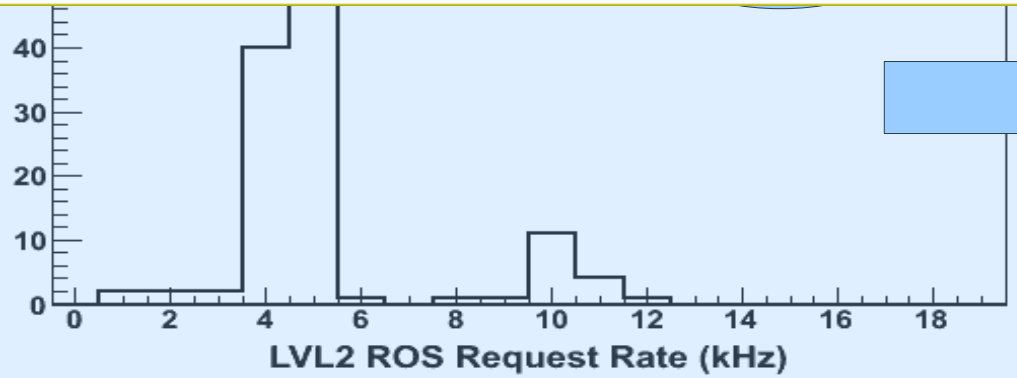
LVL2 Traffic 2.3 GB/s
EB Traffic 3.6 GB/s ↓

Closer look to the ROSes

LVL2 Average Collection Time/ROS/RoI



- LVL2 Collection Time always smaller than 1ms
  - Far from any ROS limit



LVL2 Traffic 2.3 GB/s

EB Traffic 3.6 GB/s ↓

Conclusions

- 1/3 of the ATLAS Event Builder nodes are installed and tested
- ATLAS Event builder is based on a pull protocol
  - Data Flow Manager (DFM) receives triggers and load balance the building farm
  - Event Builder application (SFI) requests data to the ROS, handles packet losses and traffic shaping, serves complete events to the Event Filter and to monitoring applications
- Extended Event Building tests exploiting HLT nodes
  - we exceed the required bandwidth with 2/3 of the building nodes
  - 10% degradation expected sending data to the Event Filter
- Successfully tested a multi-core machine running two SFI applications (Double SFI approach)
  - 15% degradation running two applications on a single node instead of two
  - still able to provide more than needed throughput per application
- Initial test of Event Builder performance including the LVL2 traffic did not expose major system limits
  - not yet able to reach the final load on the ReadOut System and the Event Builder mostly because of the limited available hardware
  - system to be extended in Spring 2008

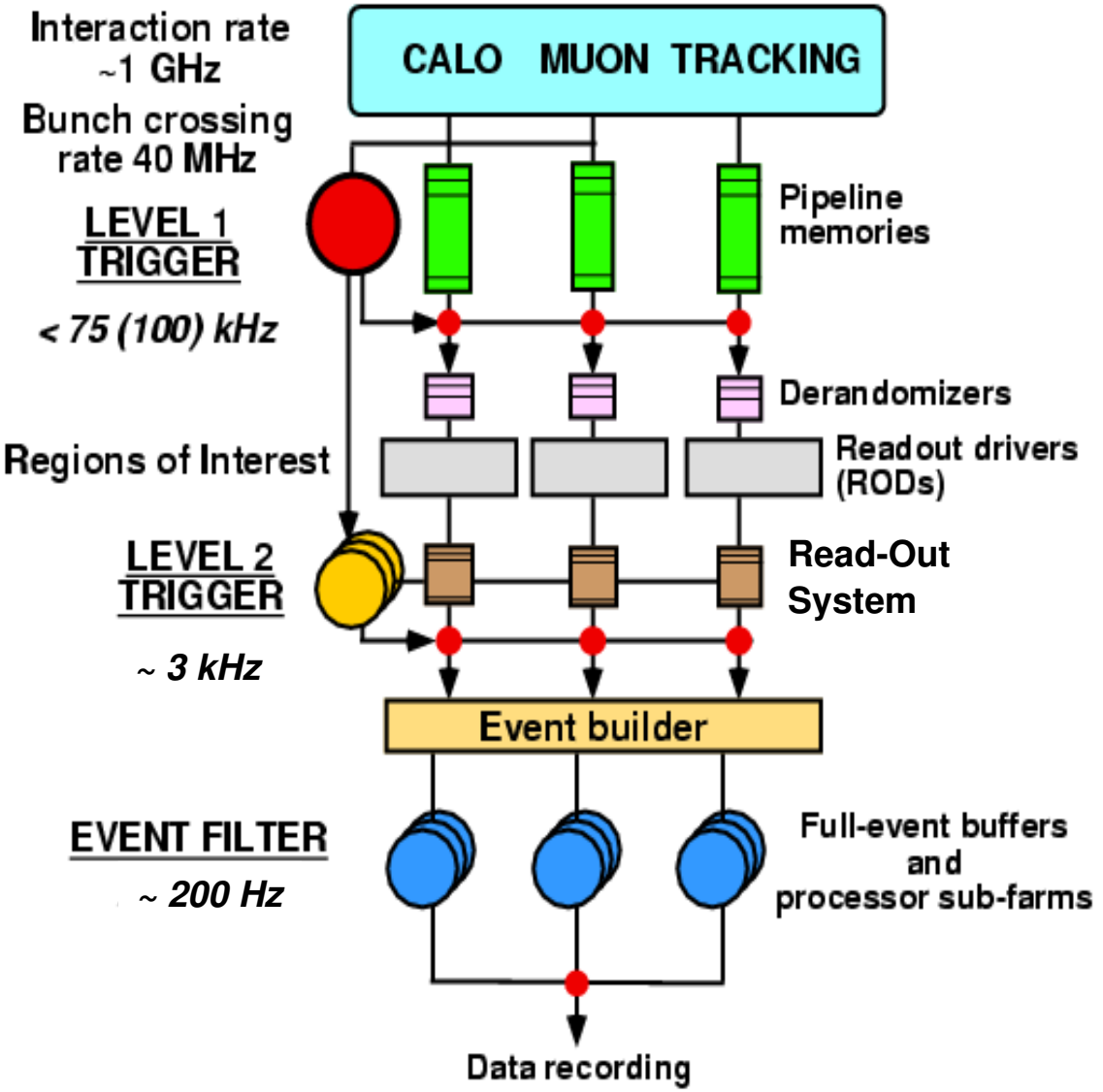
Backup Slides

Data Acquisition Strategy

- Based on **three trigger levels**
  - LVL1 hardware trigger

High Level Trigger (HLT)

- LVL2: 500 1U PC farm
  - Reconstruction within Region of Interest (RoI) defined by LVL1
- EF: 1900 1U PC farm
  - Complete event reconstruction

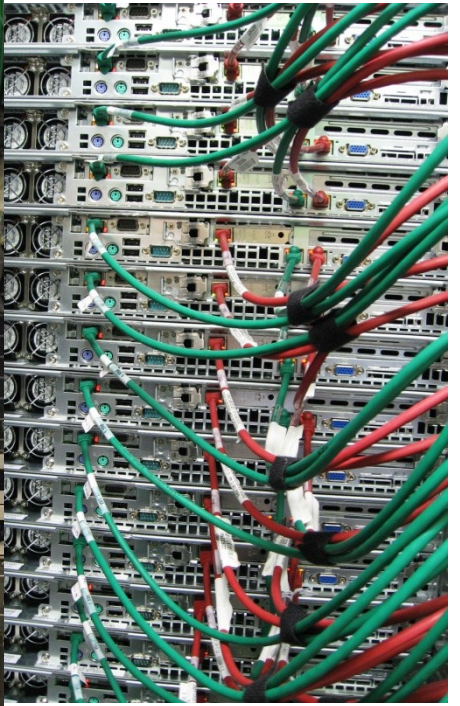


ReadOut System (ROS)

- **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

SFIs

- **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



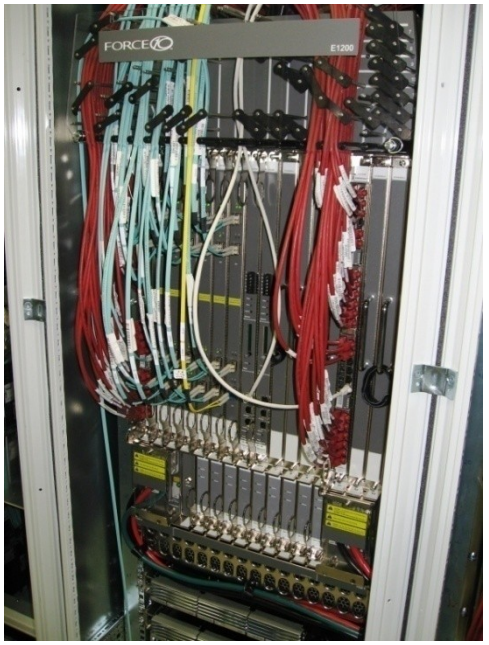
DFMs

- **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

HLT

- **124 HLT PCs installed**
- 1U, 19" rack mountable PC
- CPU: Intel Xeon E5320 quad-core 1.86GHz
  - SMP kernel
- RAM: 8 GB
- Network:
  - **2 GB onboard**
    1 used for control network
    1 used for data-in
  - 1 dedicated IPMI port
- Network booted
- Local hard disk
- Remote management via IPMI

Networking

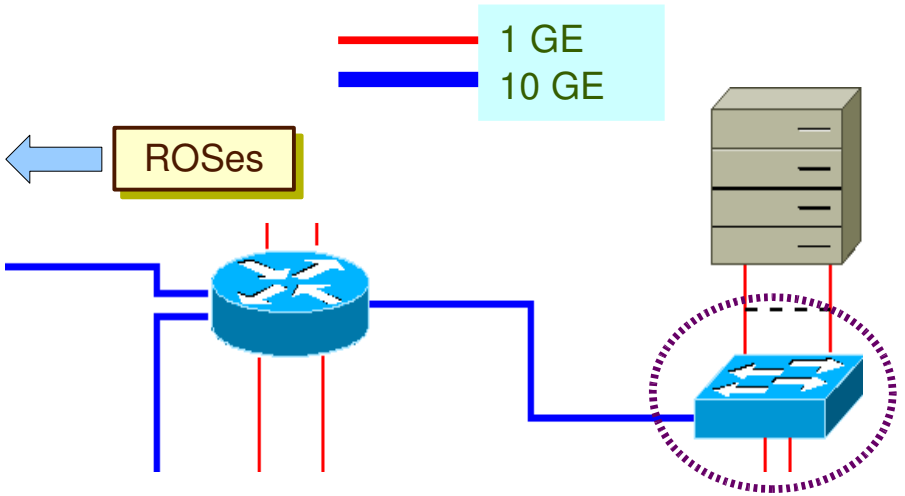


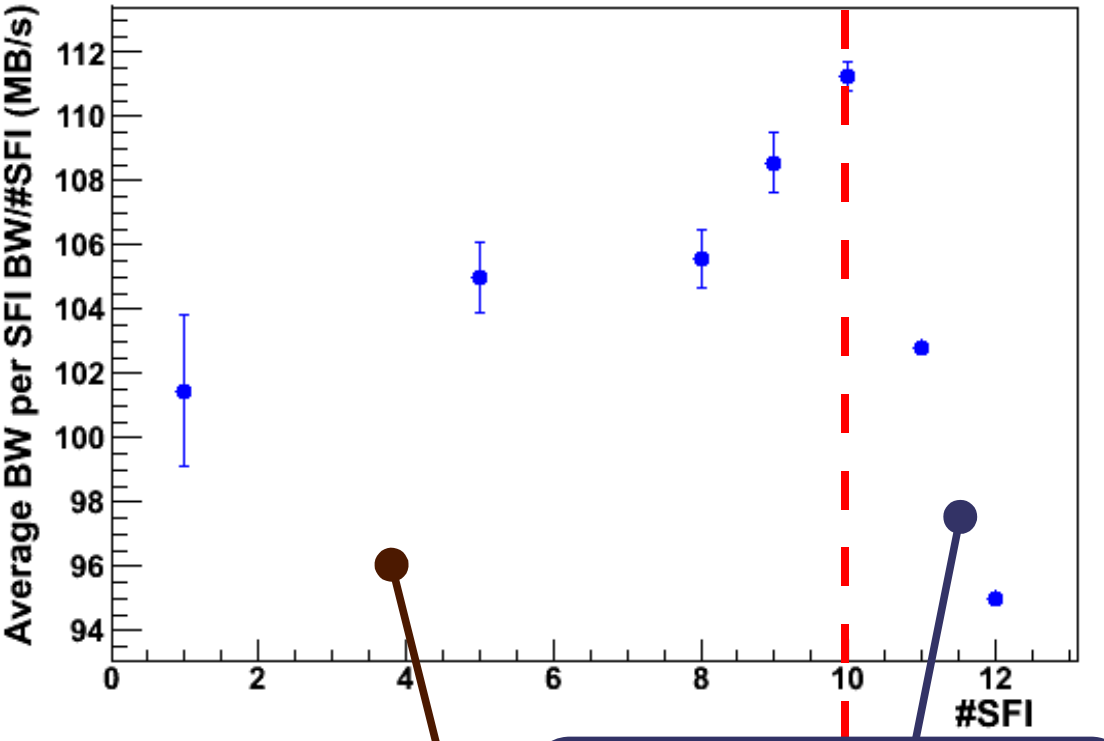
- **Force10 E1200**
  - 6 blades x 4 optical 10GE ports
  - 2 blades x48 copper GE ports
  - Up to 14 blades 1260 GE ports total 672 GE ports @ line speed
- **Data network**
  - Event builder traffic
  - LVL2 traffic

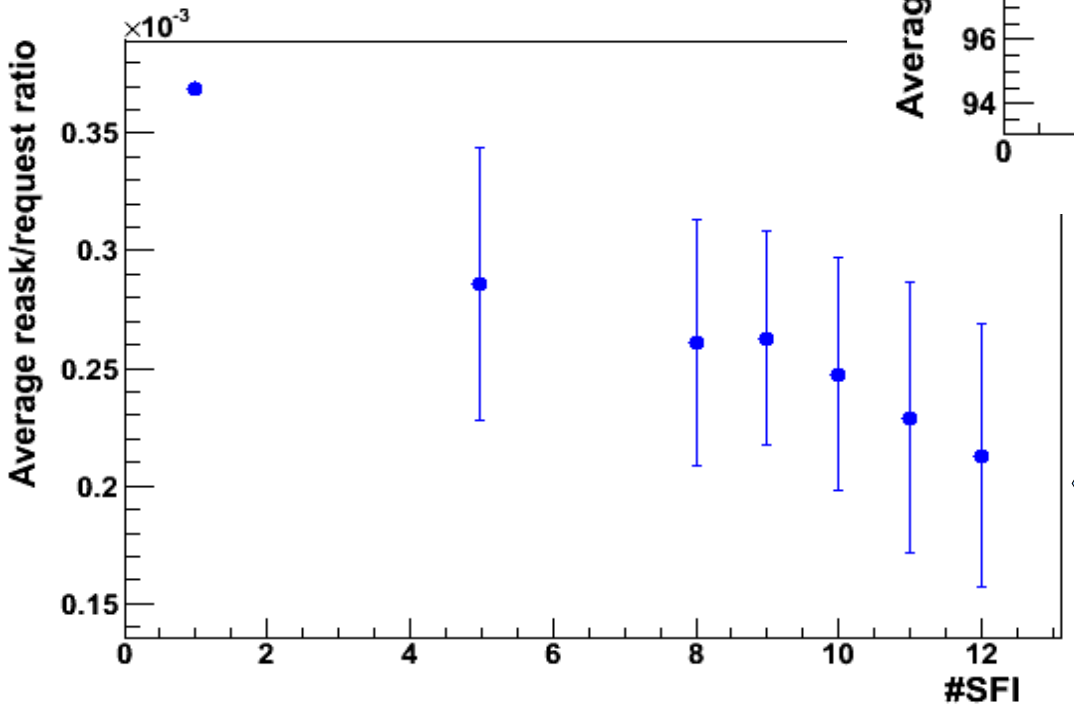
- **Force10 E600**
  - Up to 7 blades 630 GE ports total 336 GE ports @ line speed
- **Data network**
  - To Event Filter

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

HTL nodes as SFIs





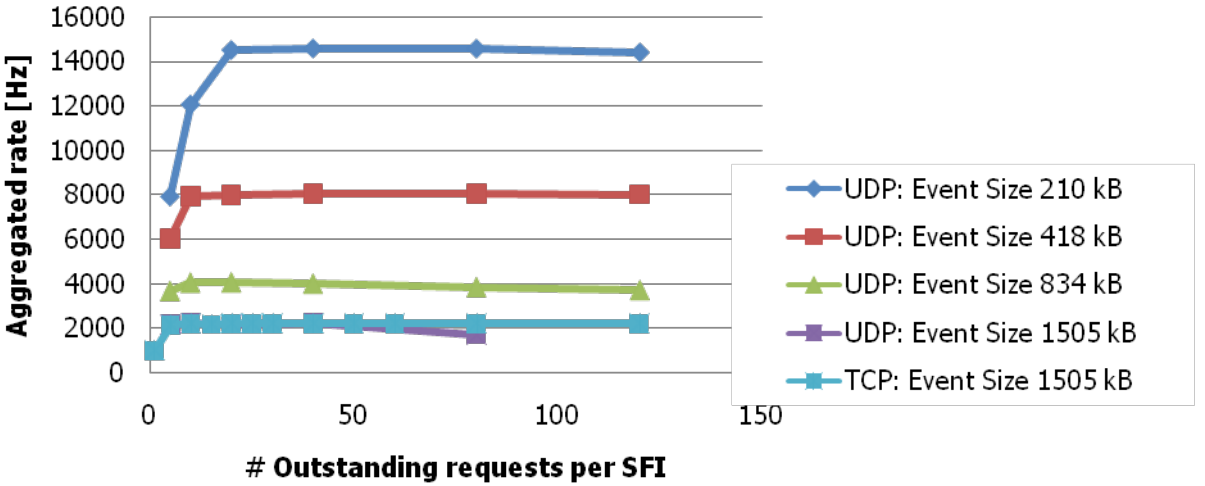


10GE link saturation

10GE link funneling into #SFI 1GE links

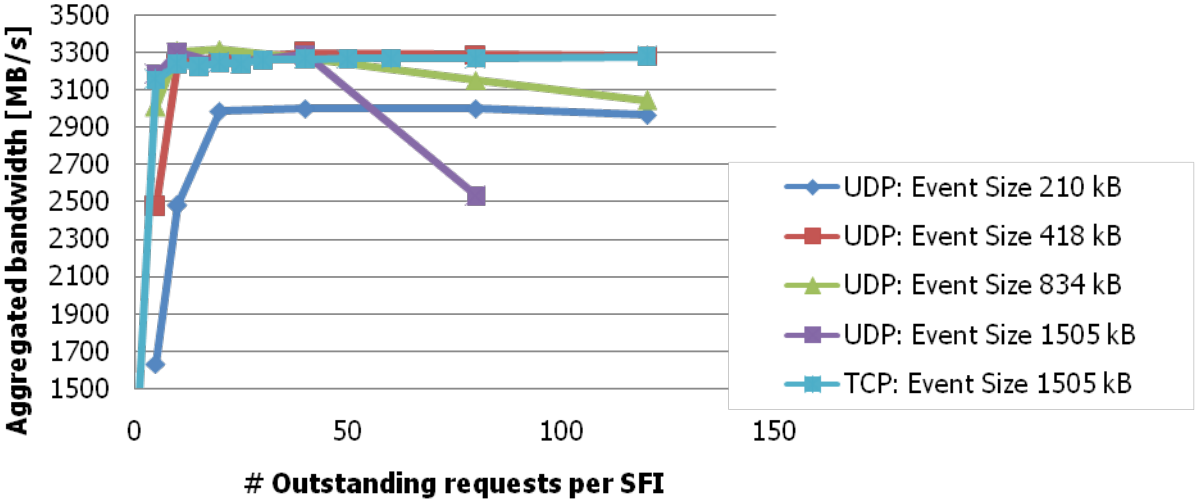
Traffic shaping

- **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



LVL2 request pattern