

*The Base line DataFlow system
of the ATLAS Trigger & DAQ*

Jos Vermeulen
NIKHEF

*On behalf of the ATLAS Trigger/DAQ
DataFlow group*

M. Abolins^a, A. Dos Anjos^b, M. Barisonzi^{c,d}, H.P. Beck^e, M. Beretta^f, R. Blair^g,
 J. Bogaerts^h, H. Boterenbrood^c, D. Botterillⁱ, M. Ciobotaru^h, E. Palencia Cortezon^h,
 R. Cranfield^j, G. Crone^j, J. Dawson^g, B. DiGirolamo^h, R. Dobinson^h, Y. Ermoline^a,
 M.L. Ferrer^f, D. Francis^h, S. Gadomski^{e,k}, S. Gameiro^h, P. Golonka^h, B. Gorini^h,
 B. Green^l, M. Gruwe^h, S. Haas^h, C. Haeberli^e, Y. Hasegawa^m, R. Hauser^a,
 C. Hinkelbein^p, R. Hughes-Jonesⁿ, P. Jansweijer^c, M. Joos^h, A. Kaczmarska^o,
 E. Knezo^h, G. Kieft^c, K. Korcyl^o, A. Kugel^p, A. Lankford^q, G. Lehmann^h,
 M. LeVine^r, W. Liu^f, T. Maeno^h, M. Losada Maia^b, L. Mapelli^h, B. Martin^h,
 R. McLaren^h, C. Meirosu^h, A. Misiejuk^l, R. Mommsen^q, G. Mornacchi^h, M. Müller^p,
 Y. Nagasaka^s, K. Nakayoshi^t, I. Papadopoulos^h, J. Petersen^h,
 P. de Matos Lopes Pinto^h, D. Prigent^h, V. Perez Reale^e, J. Schlereth^g, M. Shimojima^u,
 R. Spiwoks^h, S. Stancu^h, J. Strong^l, L. Tremblet^h, J. Vermeulen^c, P. Werner^h,
 F. Wickensⁱ, Y. Yasu^t, M. Yu^p, H. Zobernig^v, M. Zurek^o

a. Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan

b. Universidade Federal do Rio de Janeiro, COPPE/EE, Rio de Janeiro

c. NIKHEF, Amsterdam

d. Universiteit Twente, Enschede, Netherlands

e. Laboratory for High Energy Physics, University of Bern, Switzerland

f. Laboratori Nazionali di Frascati dell' I.N.FN., Frascati

g. Argonne National Laboratory, Argonne, Illinois

h. CERN, Geneva, Switzerland

i. Rutherford Appleton Laboratory, Chilton, Didcot

j. Department of Physics and Astronomy, University College London, London

k. On leave from Henryk Niewodniczanski Institute of Nuclear Physics, Cracow

l. Department of Physics, Royal Holloway and Bedford New College, University of London, Egham

m. Department of Physics, Faculty of Science, Shinshu University, Matsumoto

n. Department of Physics and Astronomy, University of Manchester, Manchester

o. Henryk Niewodniczanski Institute of Nuclear Physics, Cracow

p. Lehrstuhl für Informatik V, Universität Mannheim, Mannheim

q. University of California, Irvine, California

r. Brookhaven National Laboratory (BNL), Upton, New York

s. Hiroshima Institute of Technology, Hiroshima

t. KEK, High Energy Accelerator Research Organisation, Tsukuba


u. Department of Electrical Engineering, Nagasaki Institute of Applied Science, Nagasaki

v. Department of Physics, University of Wisconsin, Madison, Wisconsin

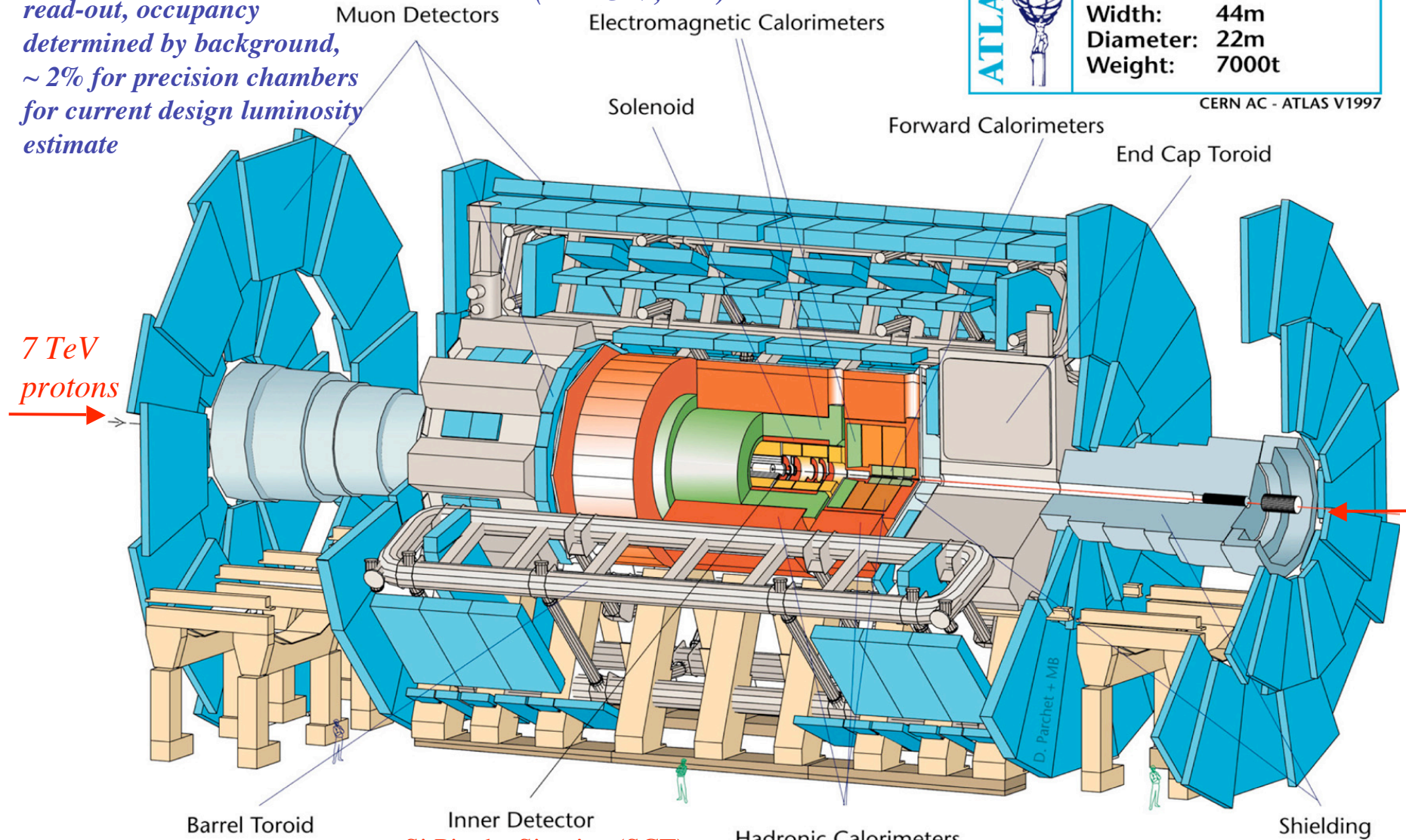
The ATLAS detector

Muon detector: zero-suppressed read-out, occupancy determined by background, ~ 2% for precision chambers for current design luminosity estimate

Calorimeters: all channels read out, very large energy depositions (> 32 GeV, rare) add some data

	Detector characteristics	
	Width:	44m
	Diameter:	22m
	Weight:	7000t

CERN AC - ATLAS V1997



Barrel Toroid

Inner Detector

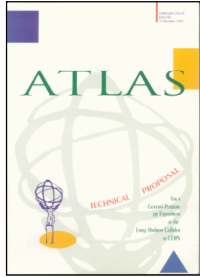
Hadronic Calorimeters

Shielding

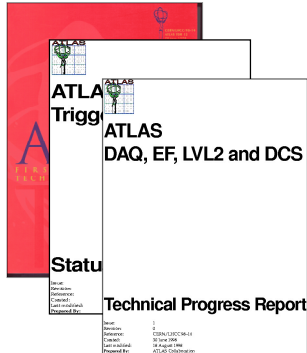
**Si Pixels, Si strips (SCT),
Transition Radiation Tracker (TRT)**

Zero-suppressed read-out, occupancy estimate design luminosity Pixels: << 1%, SCT < ~ 1%, TRT up to about 40%

TDAQ Documentation History

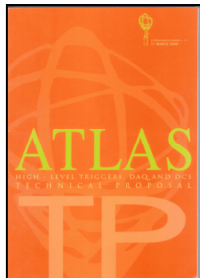


December 1994: ATLAS Technical Proposal

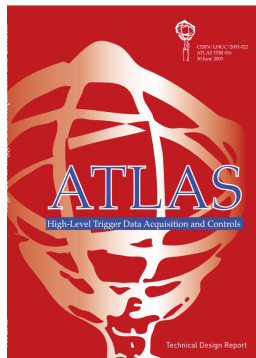


June 1998:

- Level-1 Trigger TDR
- Trigger Performance Status report
- DAQ, EF, LVL2 and DCS Technical Progress Report

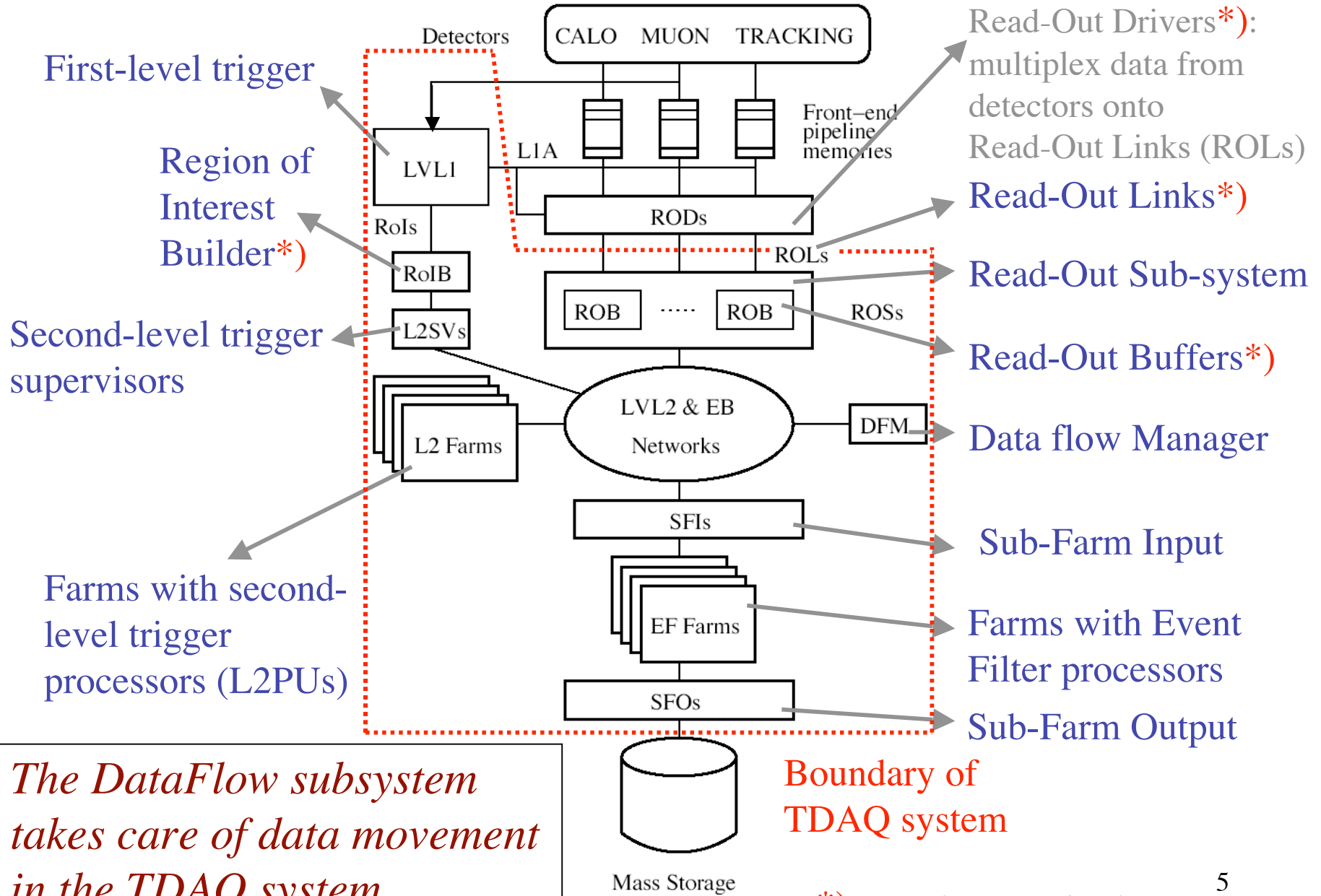


March 2000: HLT, DAQ and DCS Technical Proposal



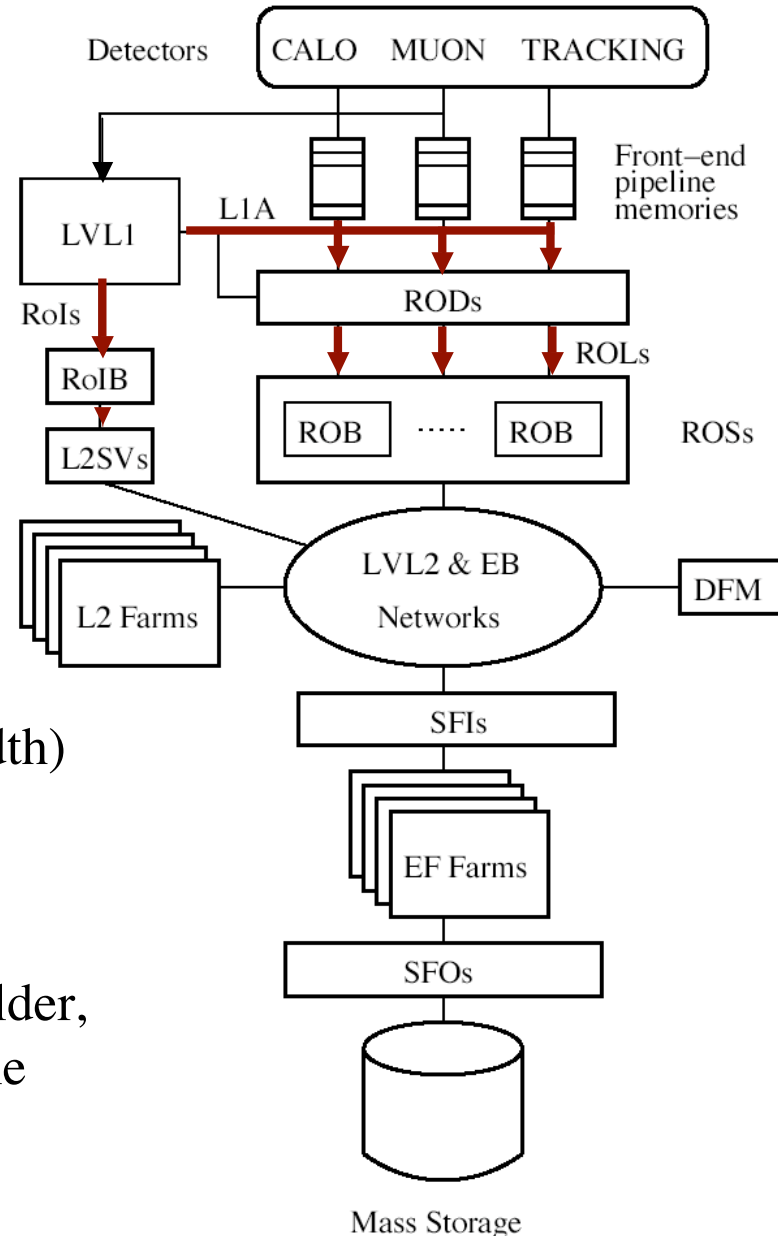
June 2003: High-Level Trigger Data Acquisition and Controls Technical Design Report

ATLAS TDAQ



Event dataflow into the ROSs and passing of RoI information

A LVL1 accept causes the front-end buffers to send event data to the RODs, which assemble event fragments and pass these via 1600 ROLs (S-LINK, optical fibers, each 160 MByte/s bandwidth) to, in the baseline design, 144 ROSs. RoI information is passed to the RoI Builder, its output is sent to one of the L2SVs.



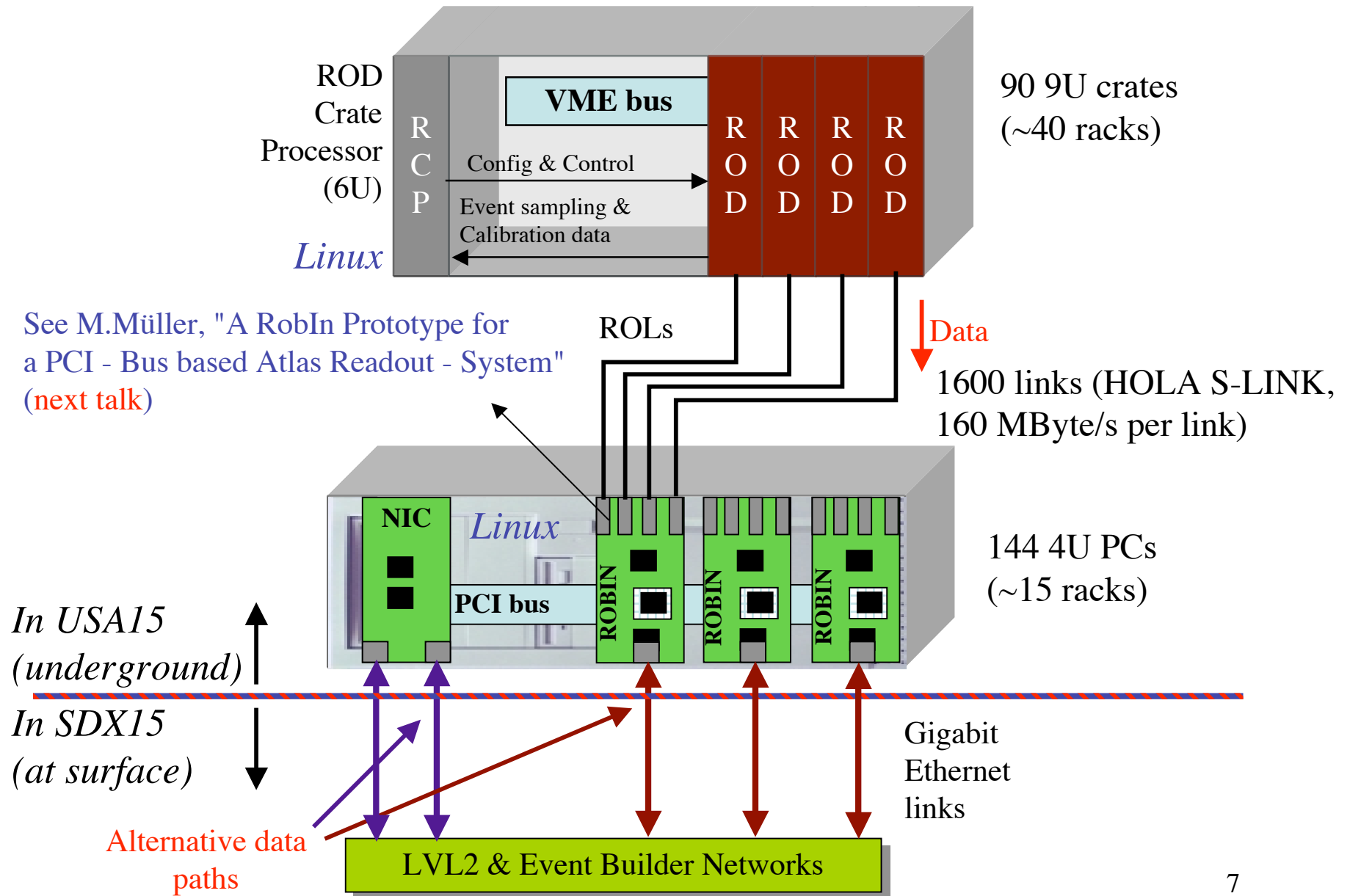
NB1: LVL1 is not part of the TDAQ DataFlow subsystem

NB2: LVL1 trigger uses data from the calorimeters and dedicated muon trigger detectors

LVL1 accept rate:
max. 75 kHz,
upgradable to 100 kHz,
nominally 25 - 40 kHz

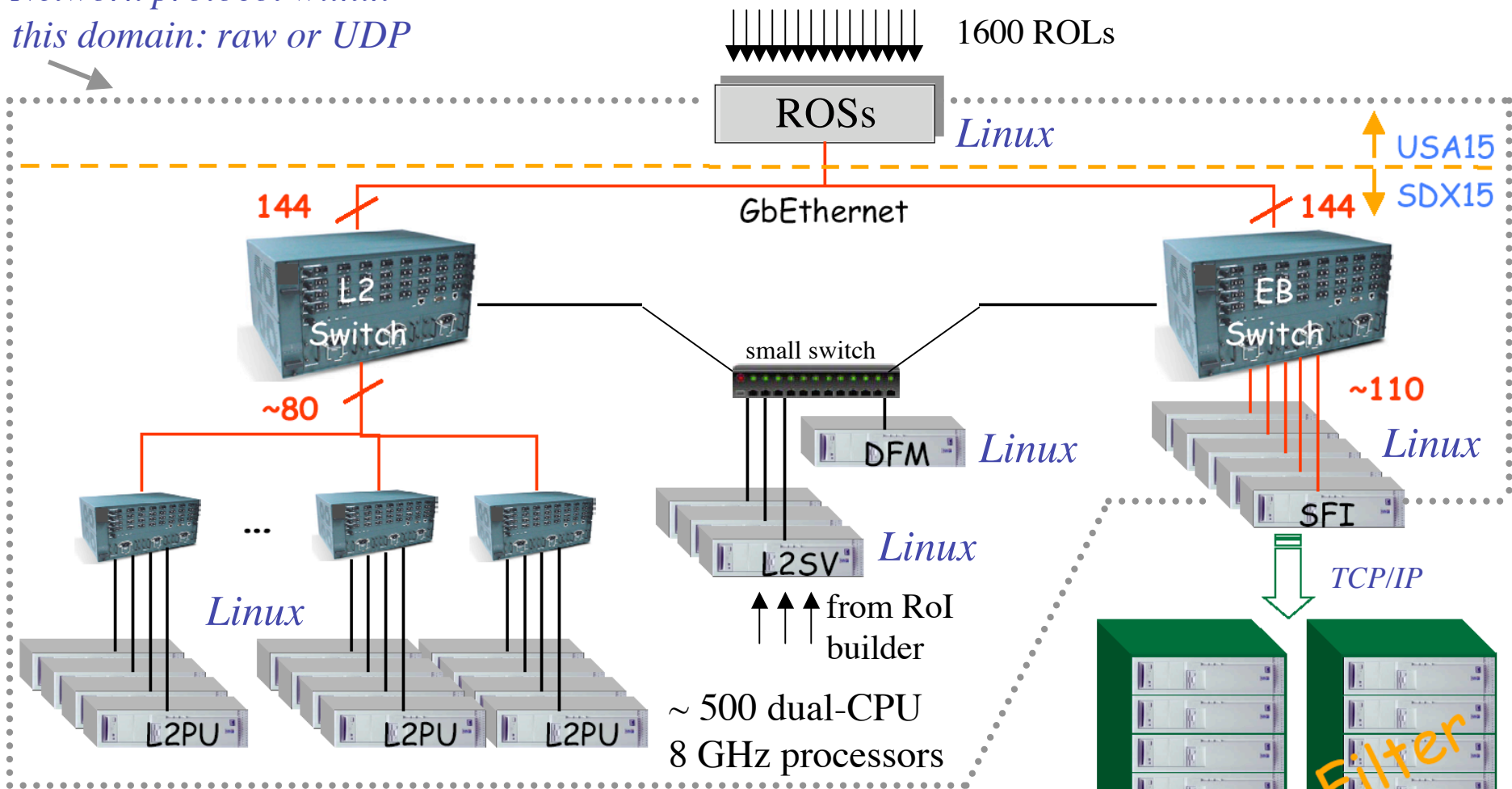
Average fragment size
per ROL < 1.6 kByte

RODs and ROSs

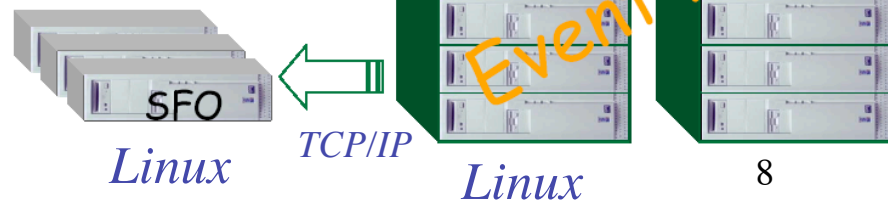


Networks, EB and HLT (LVL2 and EF)

Network protocol within this domain: raw or UDP



When individual ROBINs are connected to the network additional "concentrating" switches between the central switches and the ROBINs may be used.



*Software on Linux nodes: C++, use of POSIX threads
and of the Standard Template Library*

For detailed information see S. Gadomski, CHEP2003
"Experience with multi-threaded C++ applications
in the ATLAS DataFlow"

<http://cdsweb.cern.ch/search.py?recid=621381>
ATL-DAQ-2003-007

RoI requests

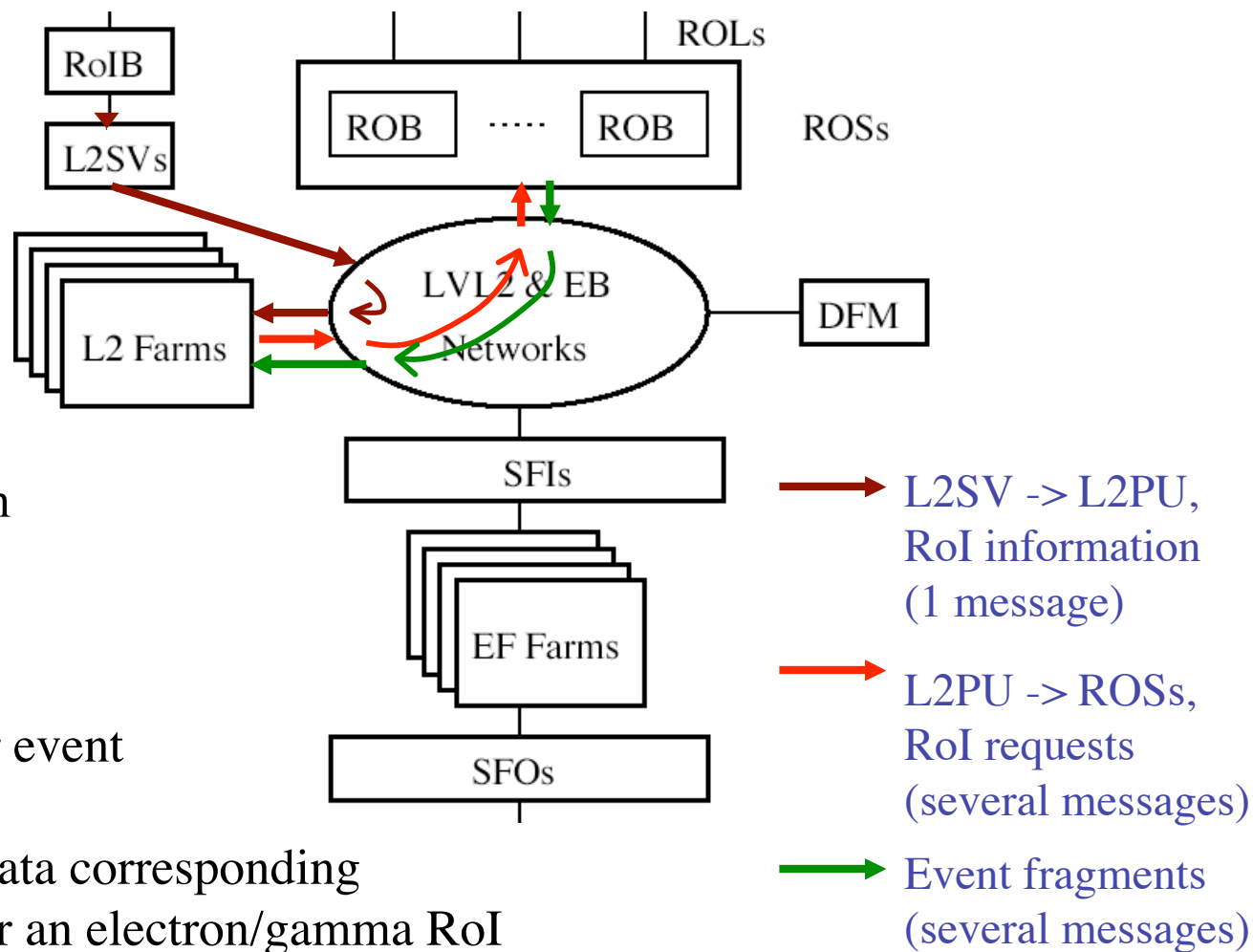
After a LVL1 accept the L2SV sends the RoI information to a L2PU.

The RoI information indicates which data has to be requested from the ROSs as input for the LVL2 selection.

On average: 1.6 RoI per event

An L2PU will request data corresponding to a RoI in steps, e.g. for an electron/gamma RoI first data from the em calorimeter, next from the hadron calorimeter and then from the inner detector.

Only a fraction of the events is accepted in each step, in the example 19% after the first step, 11 % of the original number after the second step.



RoI request rates are estimated with the "paper model"

"Paper" -> "back-of-the-envelope" calculations

In practice: C++ program (formerly spreadsheet).

Basic assumption: RoI rate does not depend on the x and y of the centre of the RoI, only on the area in x - y space associated with the RoI.

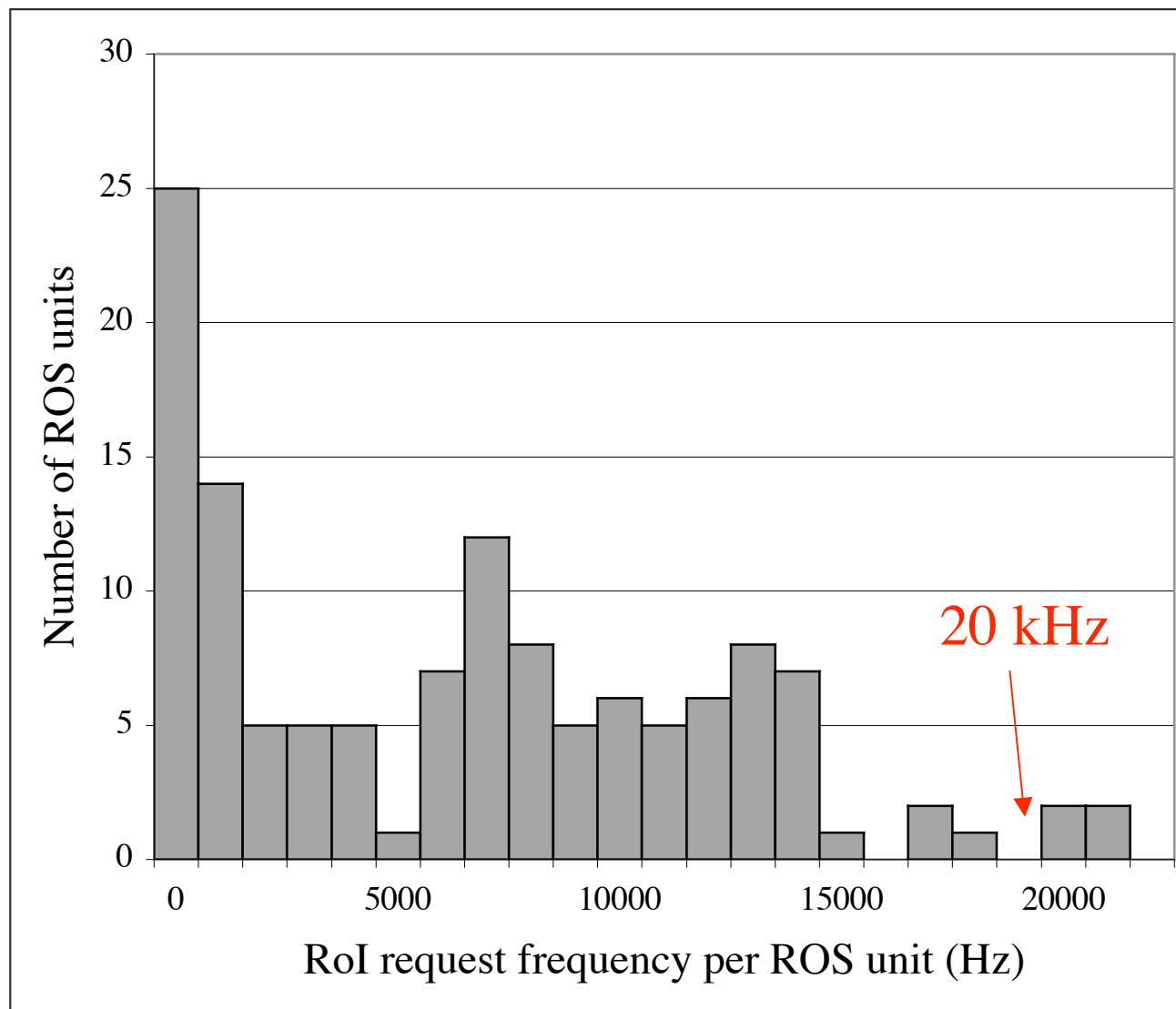
The RoI rates are obtained with a straightforward calculation using:

- the LVL1 accept rate,
- exclusive rates for the various LVL1 trigger menu items,
- the number of RoIs associated with each trigger item,
- the x - y area associated with each possible RoI location.

The request rates are then obtained using:

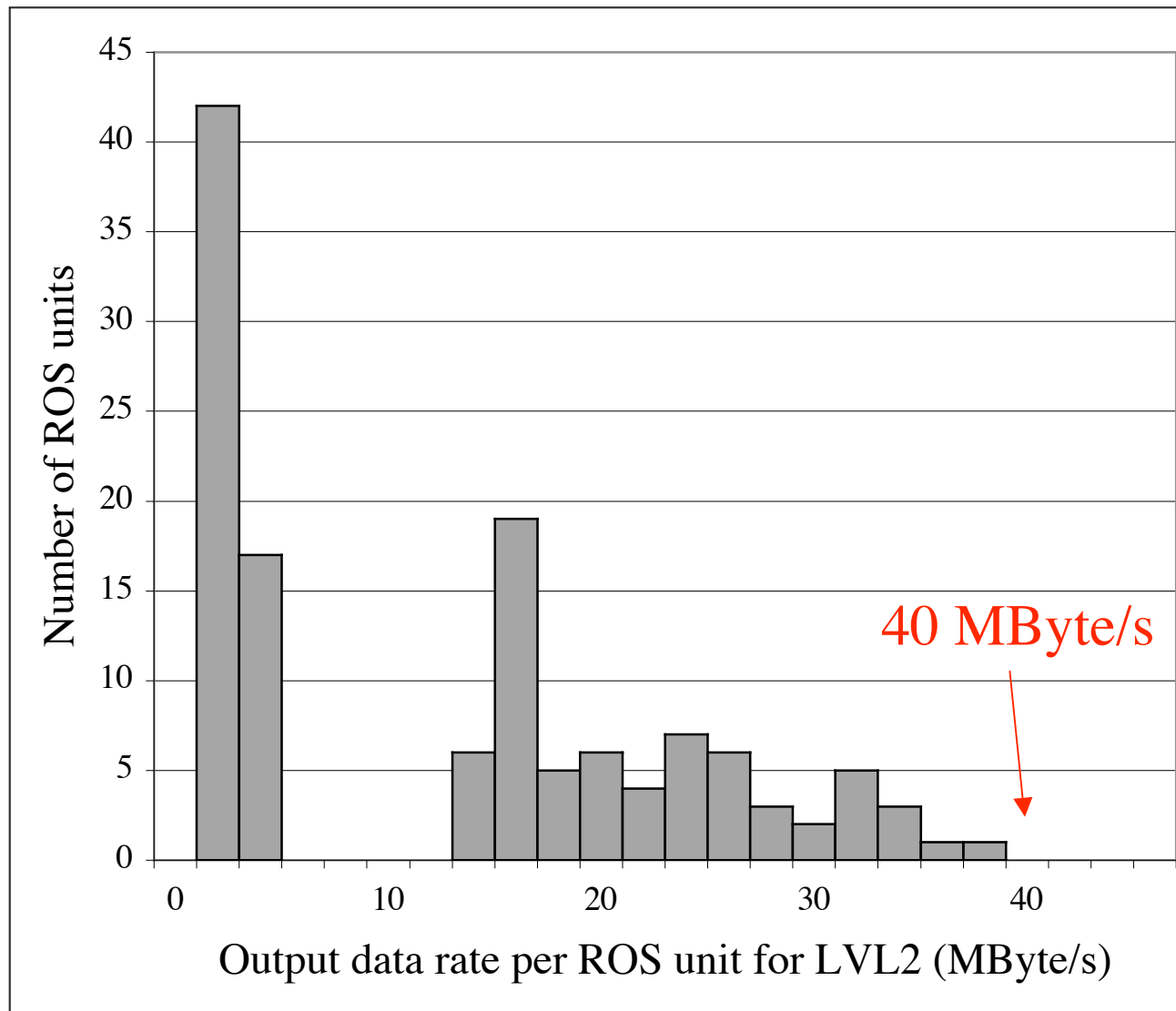
- information of the mapping of the RoIs onto the detector,
- the acceptance factors of the various LVL2 trigger steps,
- the x - y areas from which data is requested (RoI and detector dependent).

Paper model result (luminosity: $2 \cdot 10^{33}$)



LVL1 accept rate: 100 kHz

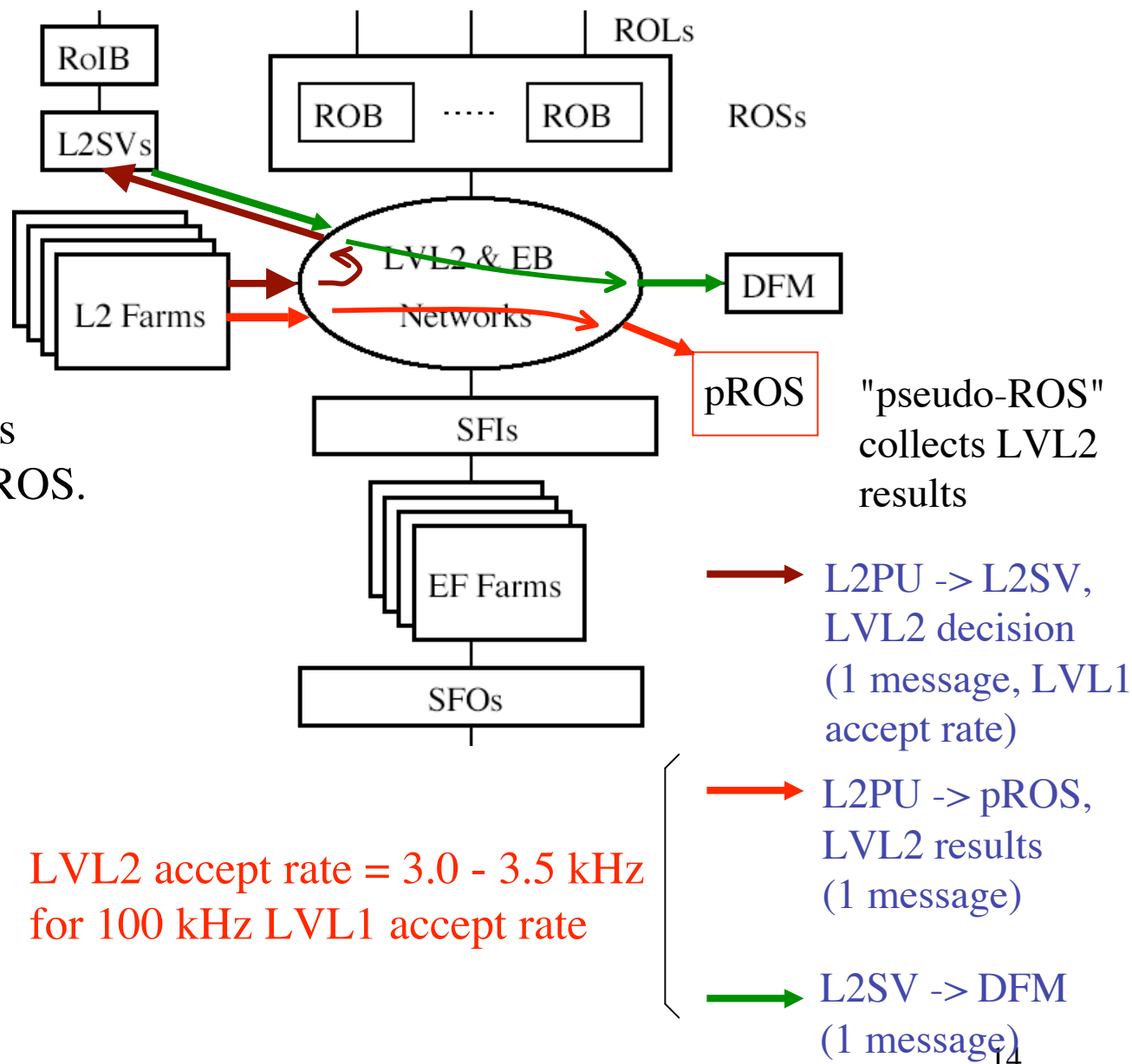
Paper model result (luminosity: $2 \cdot 10^{33}$)



LVL1 accept rate: 100 kHz

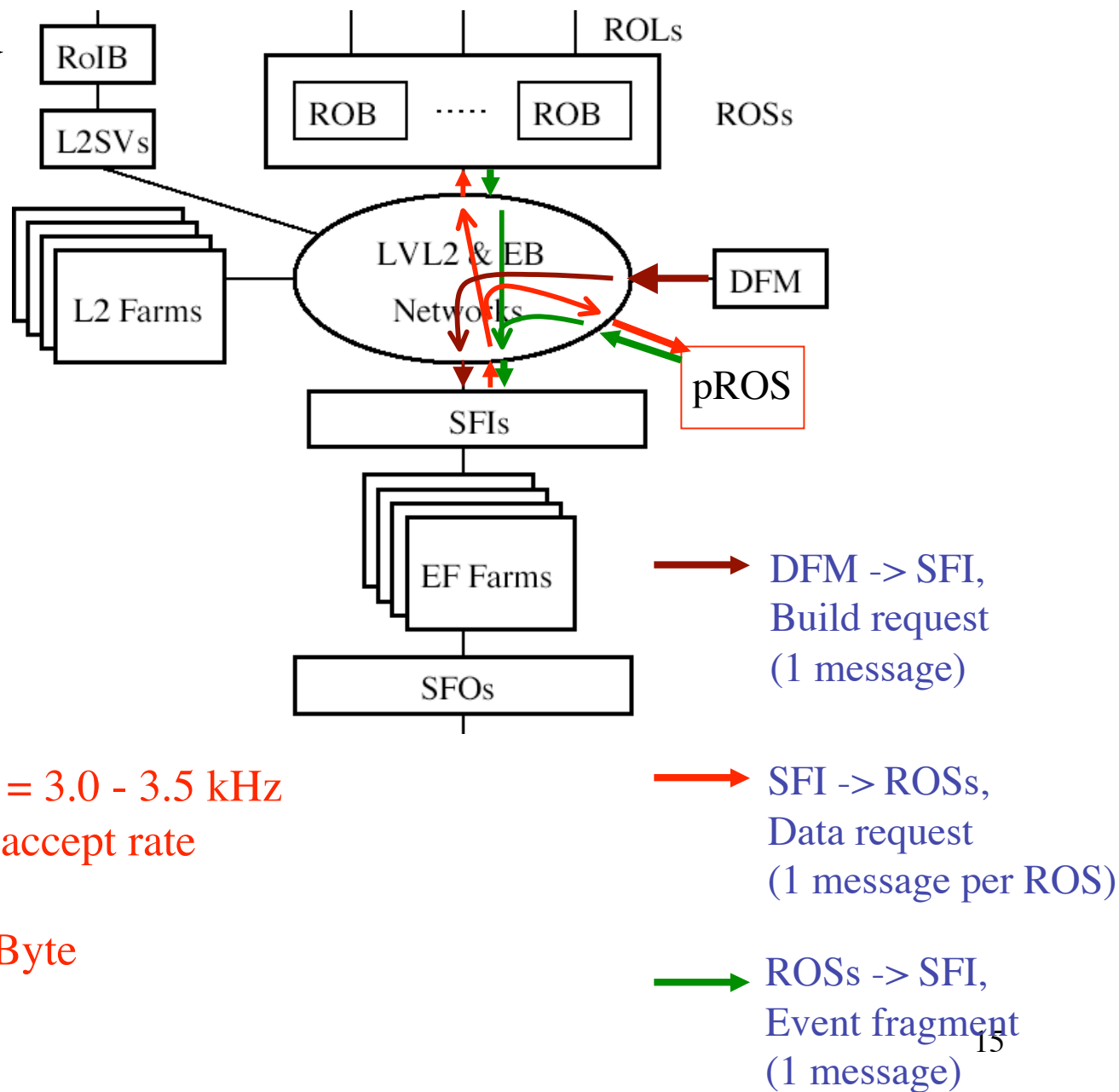
LVL2 output

After production of a decision by a LVL2 processor, the decision is communicated to the L2SV which sent the RoI request. For events accepted data produced by the trigger algorithms are also passed to the pROS.



Event building

For each event accepted by LVL2 the DFM sends a build request to an SFI. This in turn sends requests for data to the ROSs (including the pROS). The ROSs return the fragments (identified by the LVL1 id) requested.

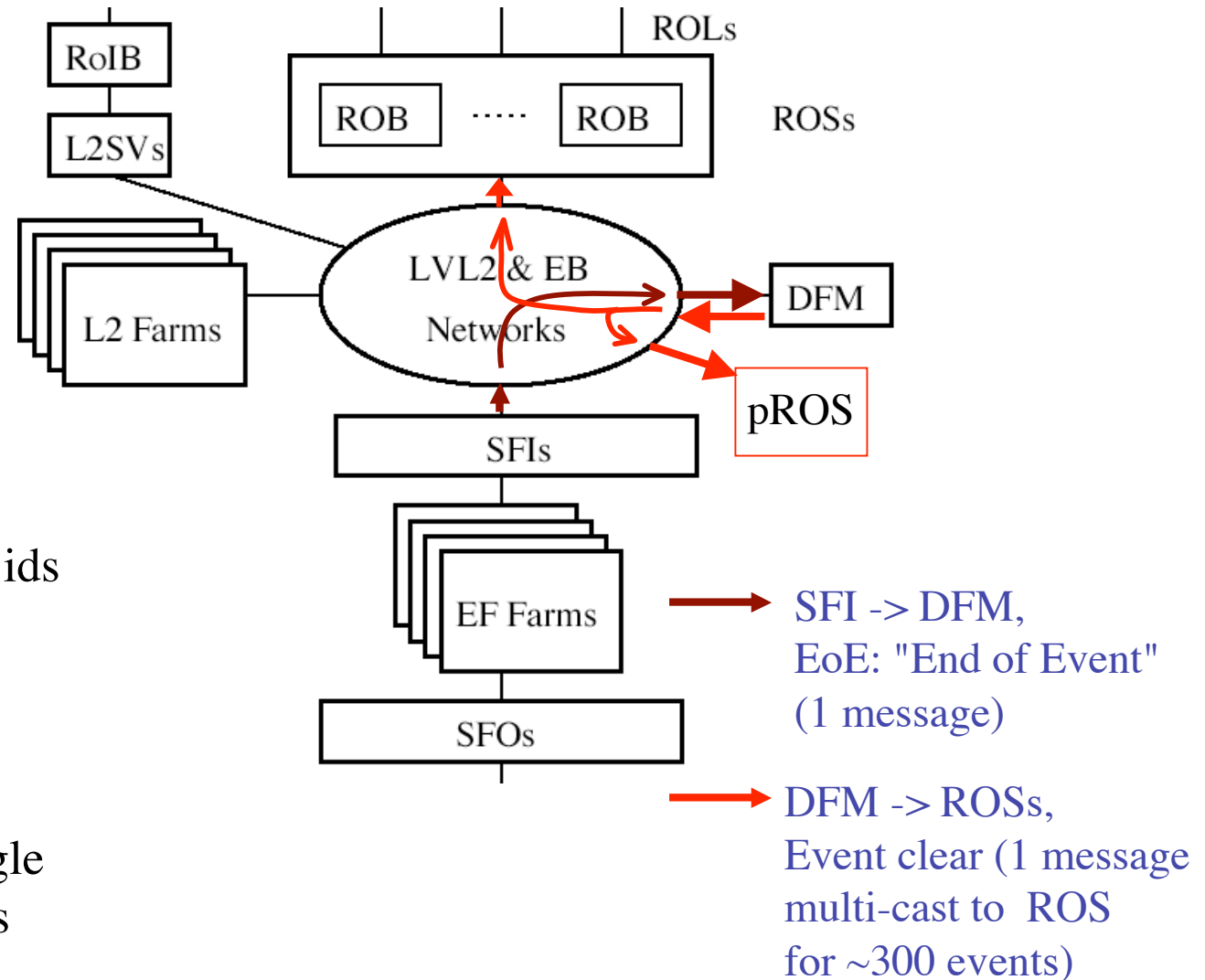


Event building rate = 3.0 - 3.5 kHz
for 100 kHz LVL1 accept rate

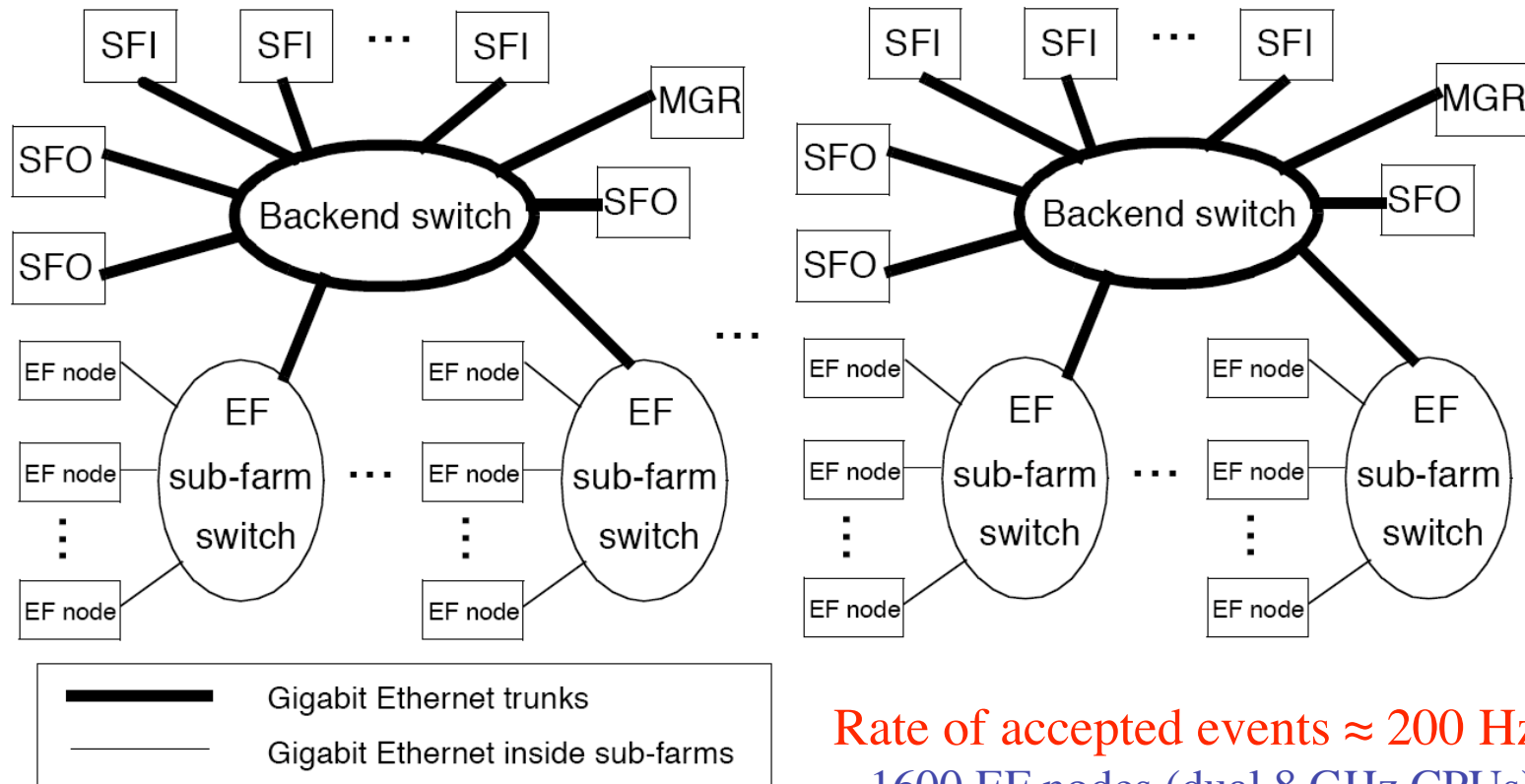
Event size ~ 1.5 MByte

Event clearing

After completion of event building an EoE (End of Event) message is sent by the SFI to the DFM. The DFM stores these and LVL2 reject messages until ~ 300 of these have been received. Event clear commands for the LVL1 ids associated with the EoE and LVL2 reject messages are then sent to the ROSs, with ~ 300 of these commands in a single message. These messages are multi-cast, and the rate is the LVL1 accept rate divided by the blocking factor (330 Hz for a LVL1 accept rate of 100 kHz).

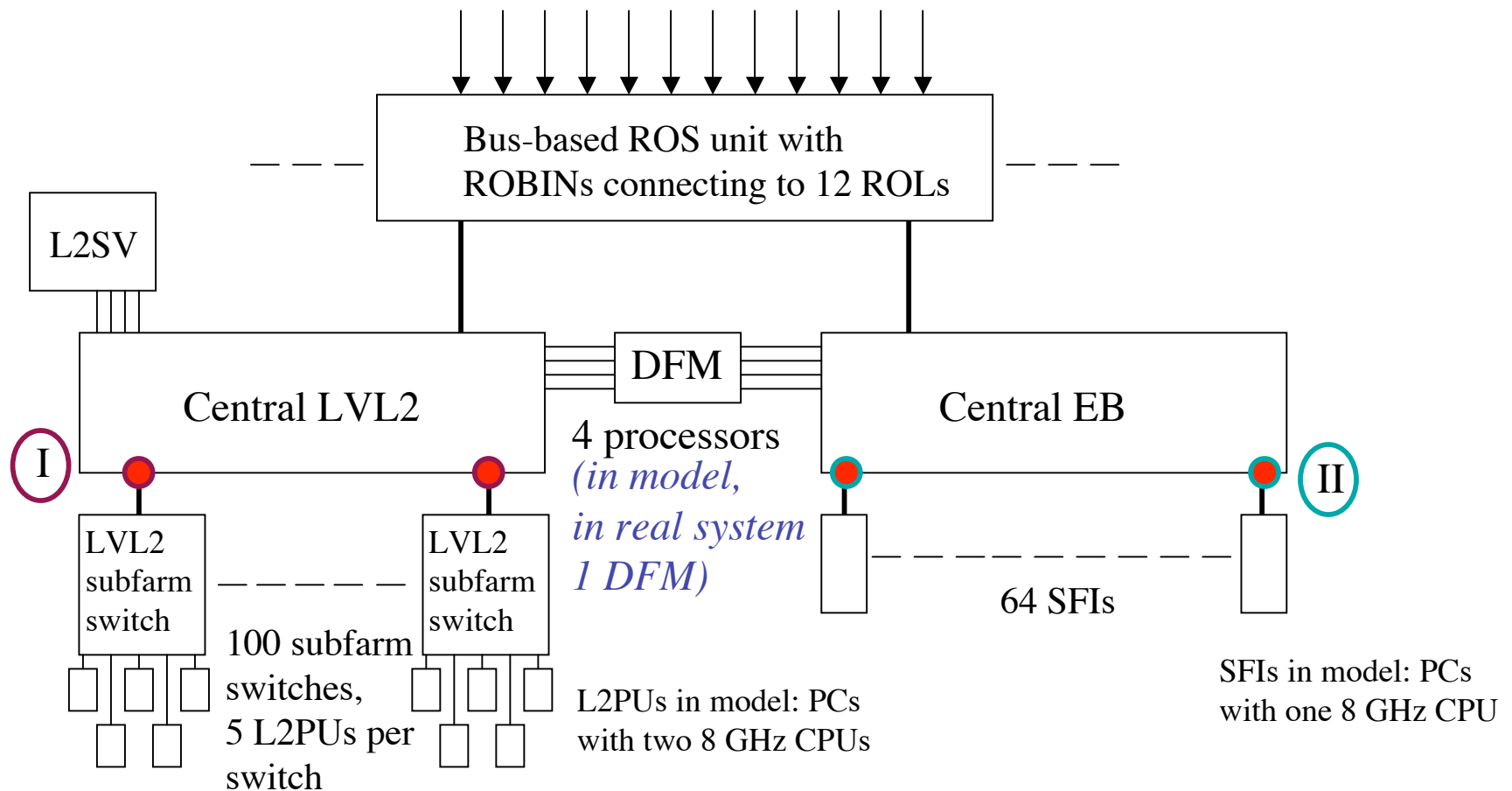


Event Filter and mass storage



After building the event it is delivered to one of the Event Filter processors (on request by these processors). A further decision is taken on acceptance or rejection. The data of accepted events are passed to the SFOs, where the events are buffered and passed to central mass storage in the CERN computer centre.

Modelling of queue formation in switches



Queues tend to form at I and II.

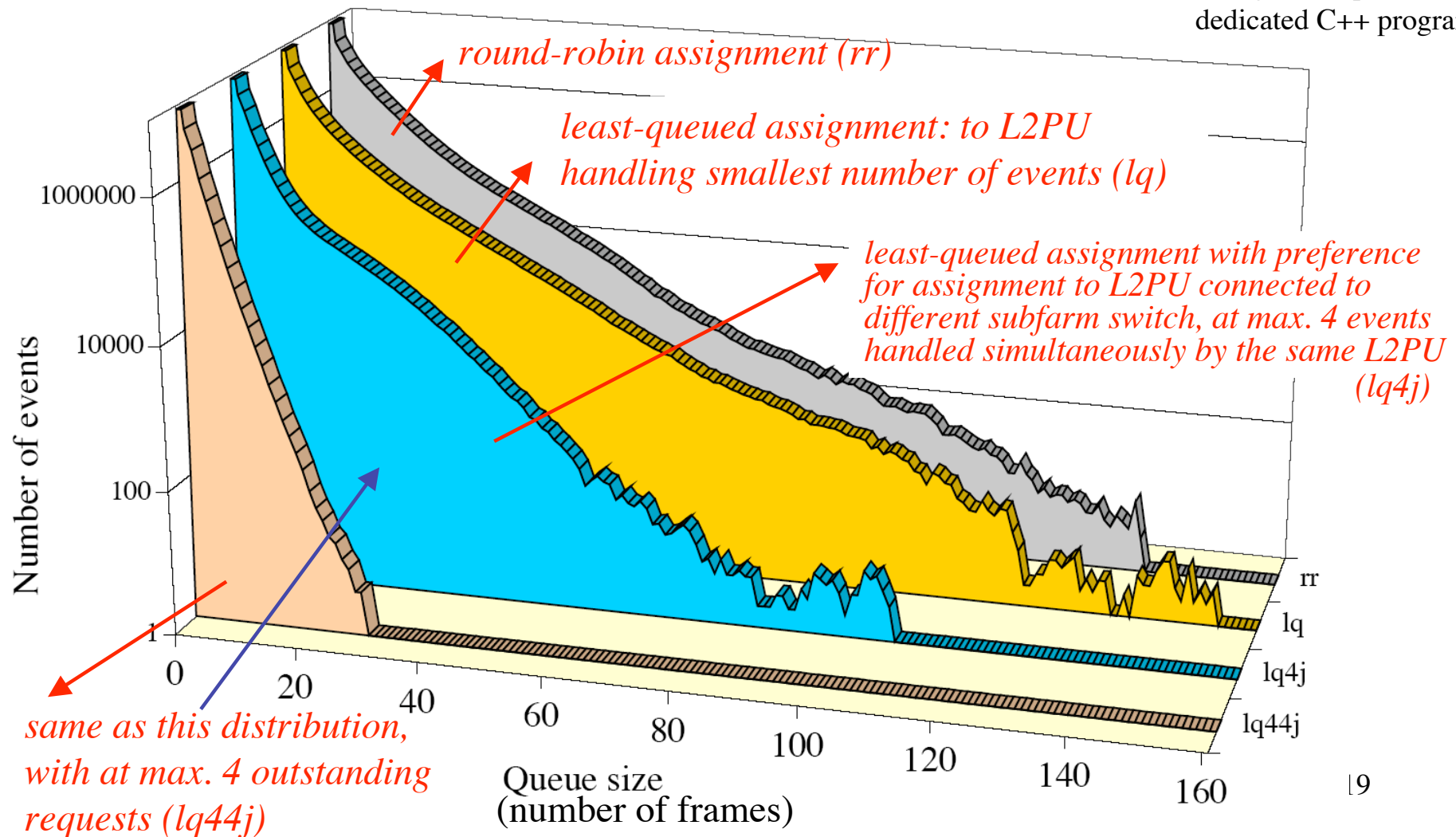
Can be controlled:

- by limiting number of events assigned simultaneously to each L2PU/SFI
- with assignment pattern of events to L2PUs/SFIs
- by limiting number of outstanding requests per L2PU/SFI

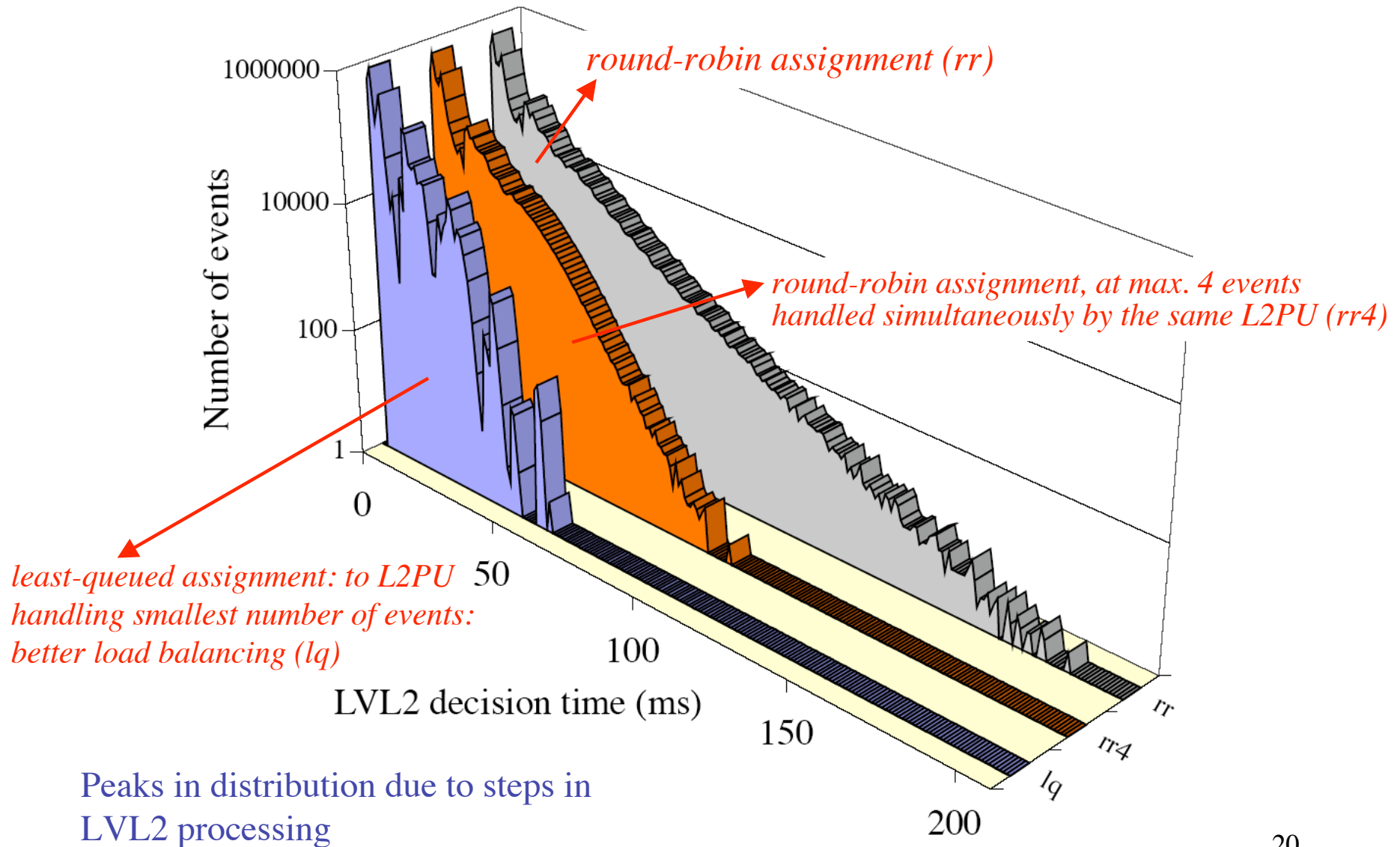
Results for point I

Obtained with discrete event simulation¹⁾, assuming use of raw Ethernet, with paper model assumptions for trigger menus, ROL mapping, acceptance factors of the different stages of LVL2 processing, 100 kHz LVL1 accept rate, design luminosity. Switches are assumed to be crossbar switches with buffers on the output ports (no flow control).

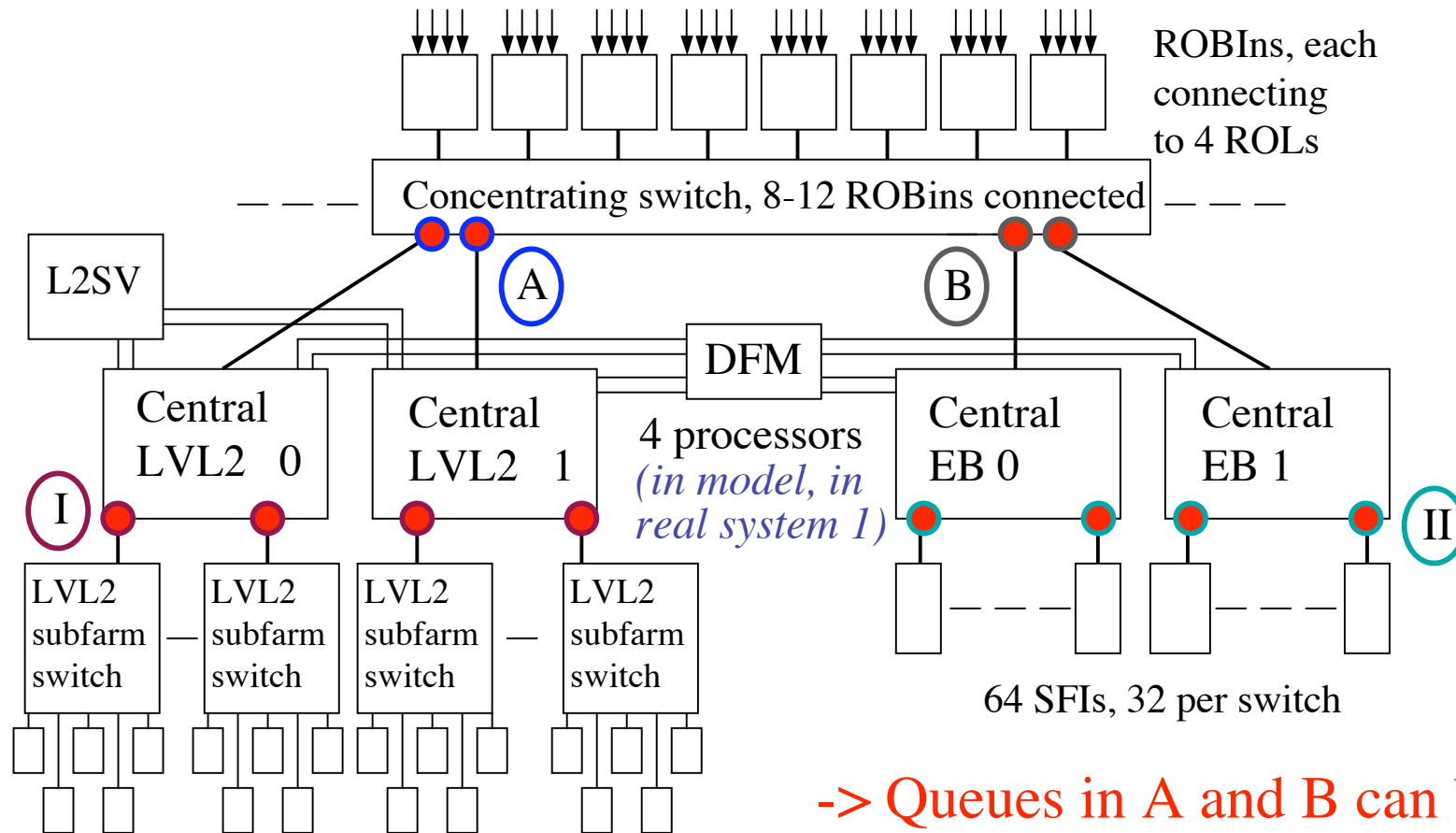
¹⁾ using simdaq, a dedicated C++ program



LVL2 decision time for same model



Modelling of system with direct connections of ROBIns to network

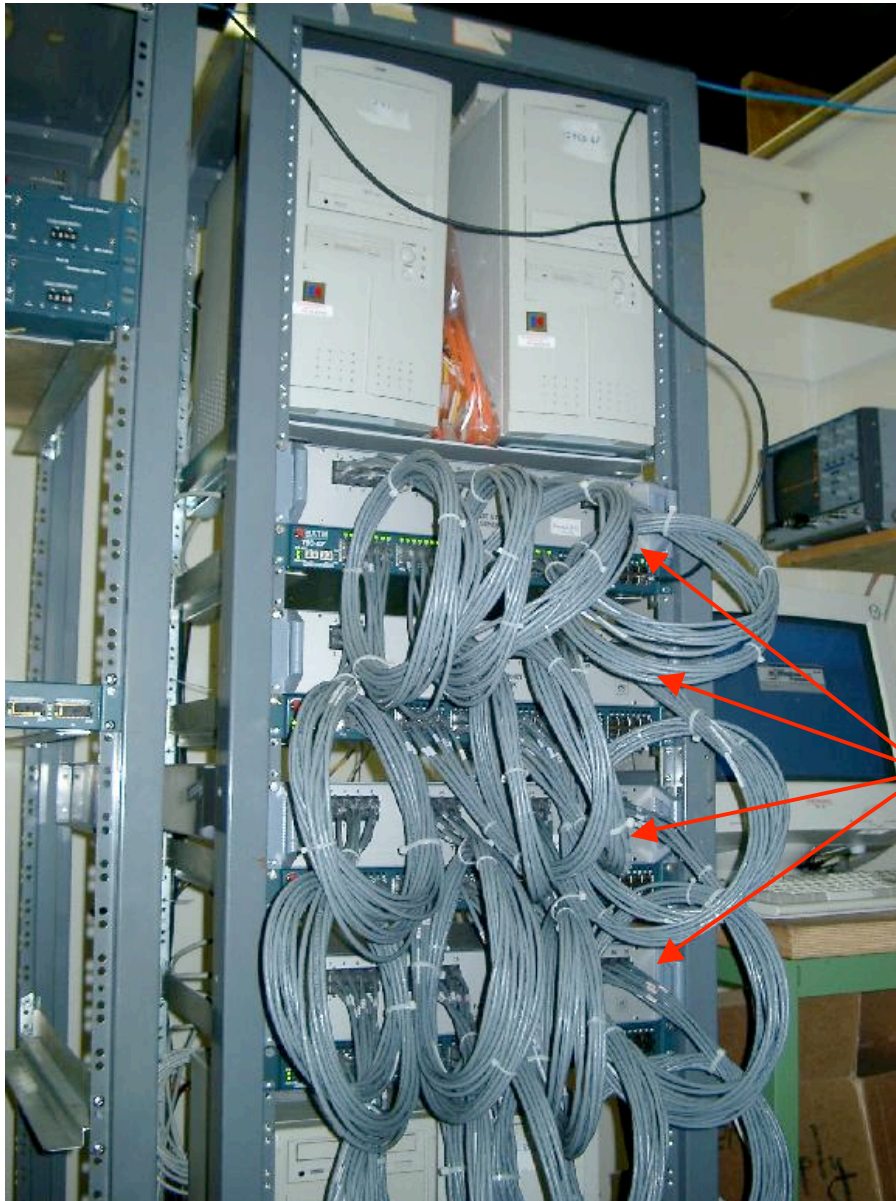


50 subfarm switches per central switch,
5 L2PUs per subfarm switch

-> Queues in A and B can be controlled with request pattern (in particular important for B)

NB: flow control can prevent buffer overflow, but may cause temporarily blocking of data transfers not affected by the buffer overflow

Testbed 1, CERN, bdg. 513

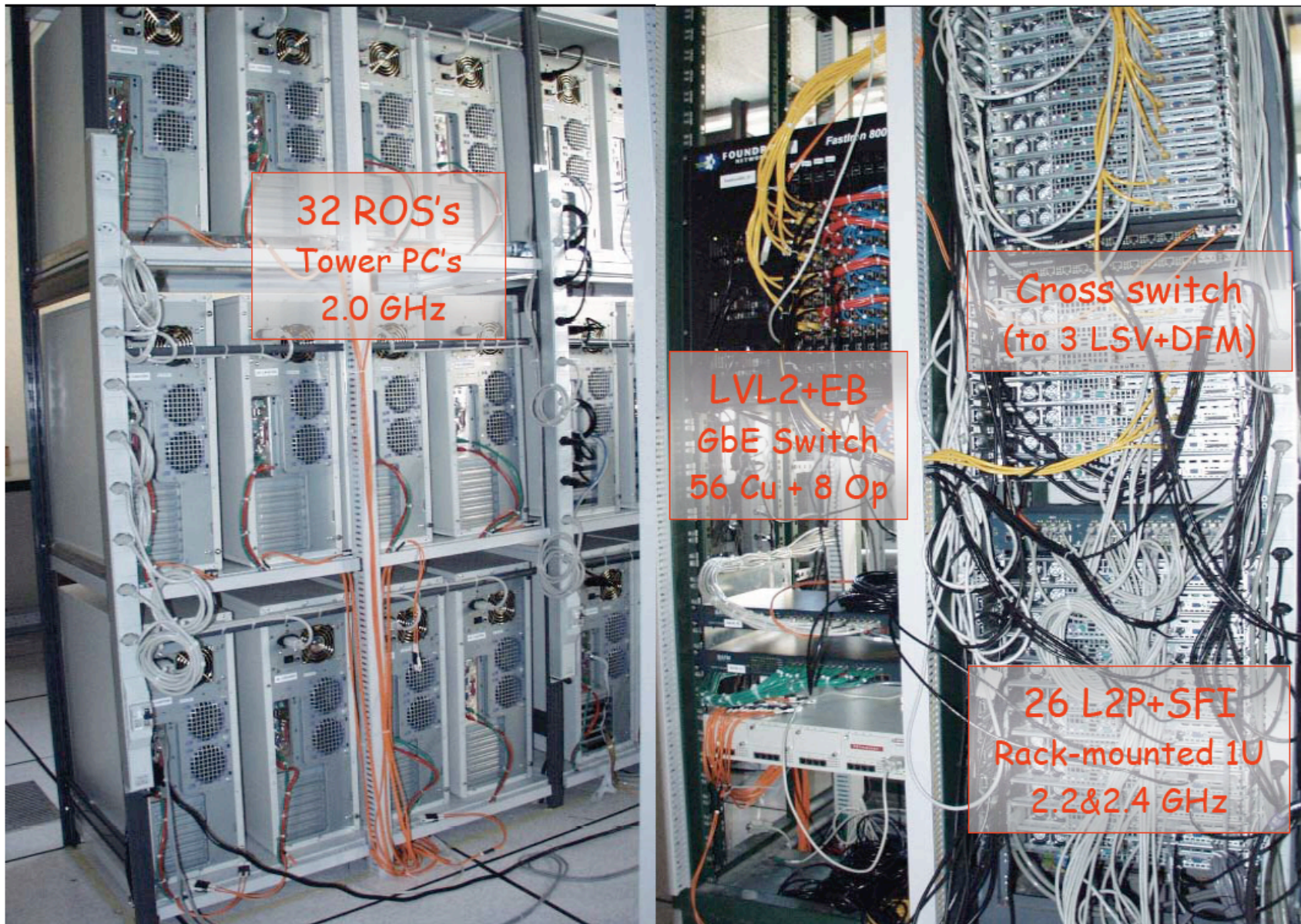


PCs in testbed: 2 - 2.4 GHz Xeon dual-CPU rack-mounted machines

128 FPGA traffic generators (4 units) each driving 32 Fast Ethernet links. Below each unit: concentrating switch (BATM T5)

Also in testbed: 8 Gigabit Ethernet traffic generators based on Alteon NICs

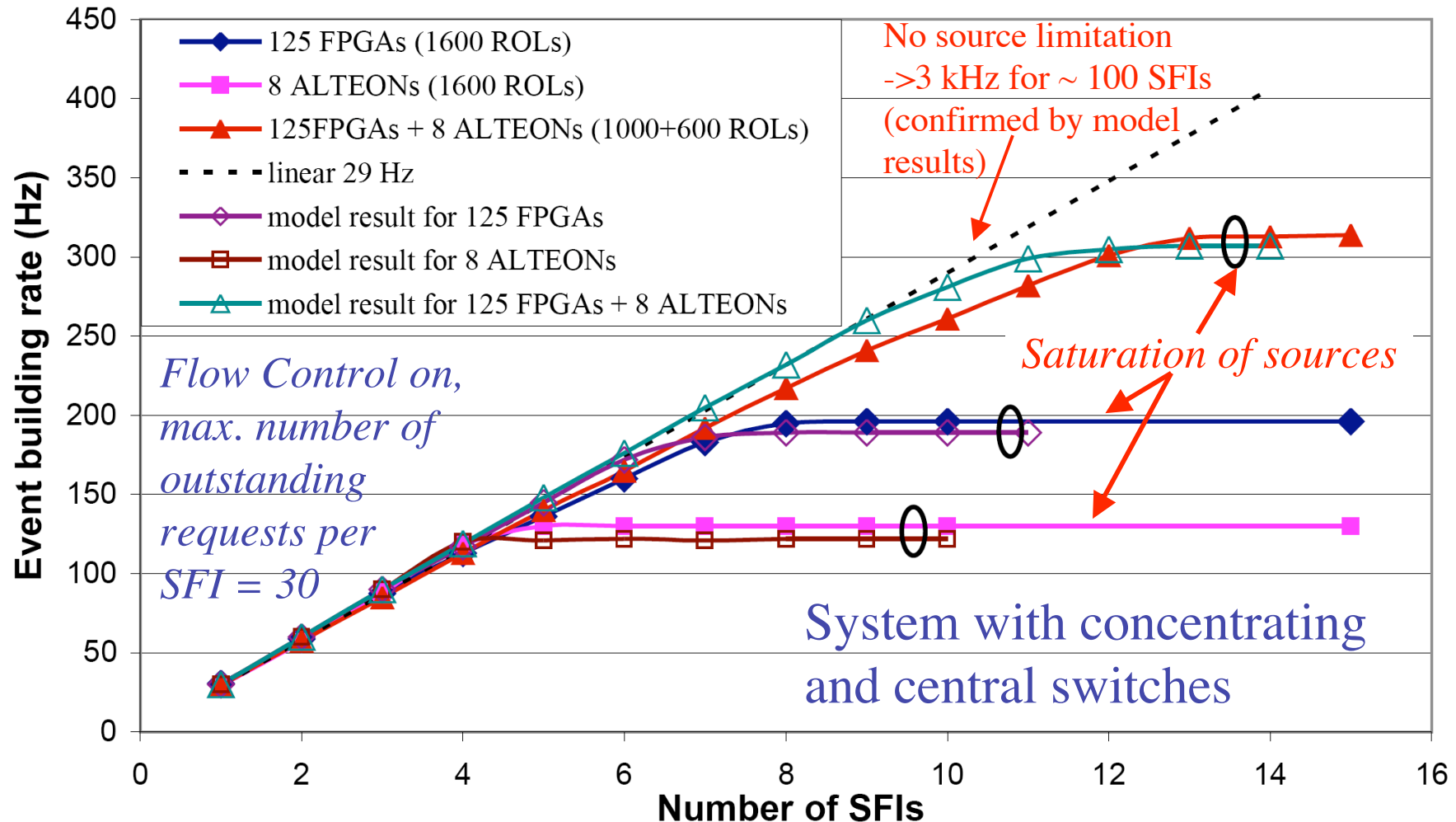
Testbed 2, CERN, bdg. 32



Linux kernel: 2.4.18 Uni-Processor (ROS) and 2.4.20, SMP

Results for event building obtained with traffic generators

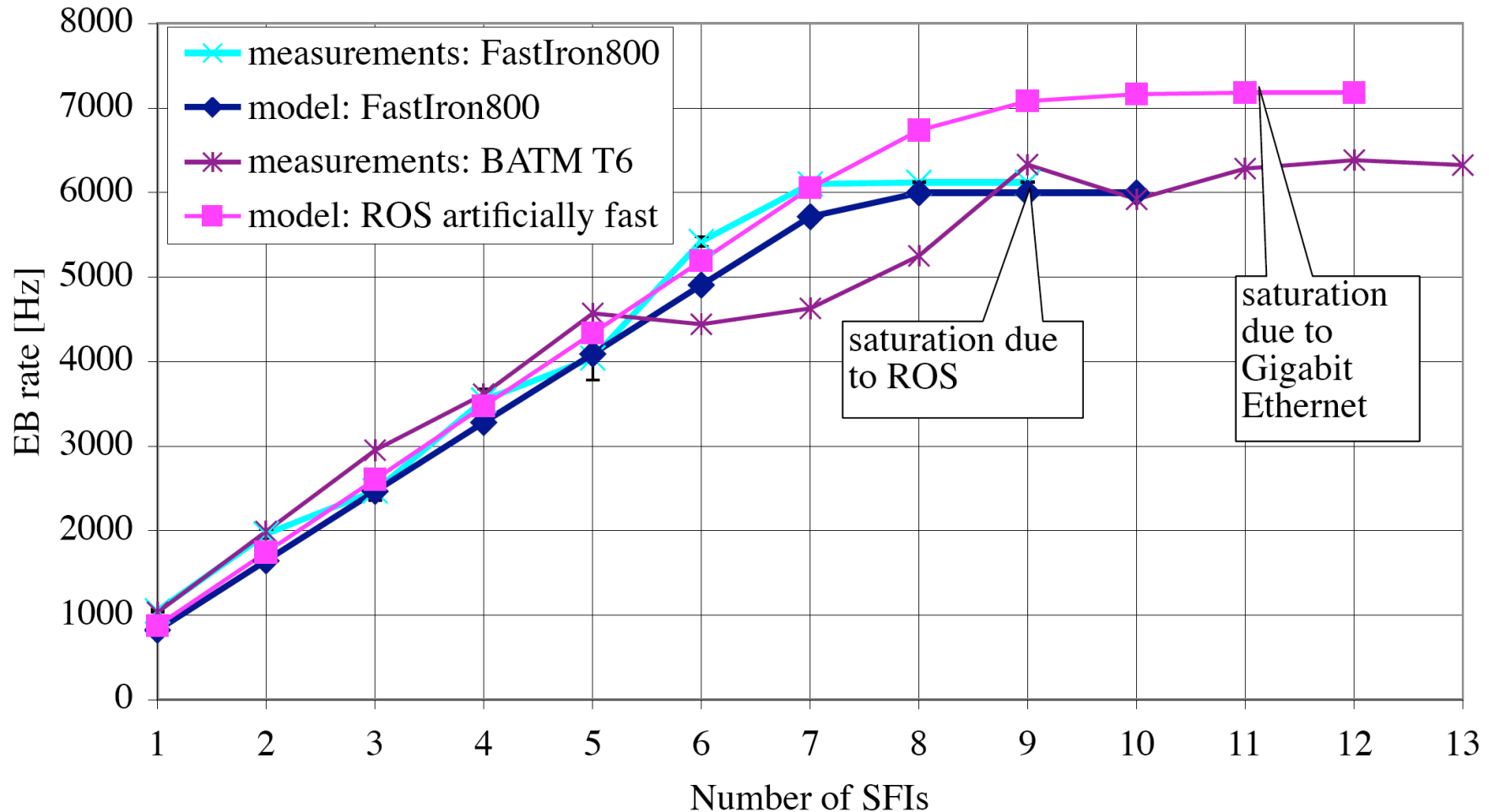
Each traffic generator emulates 8, 13, 125 or 200 data sources



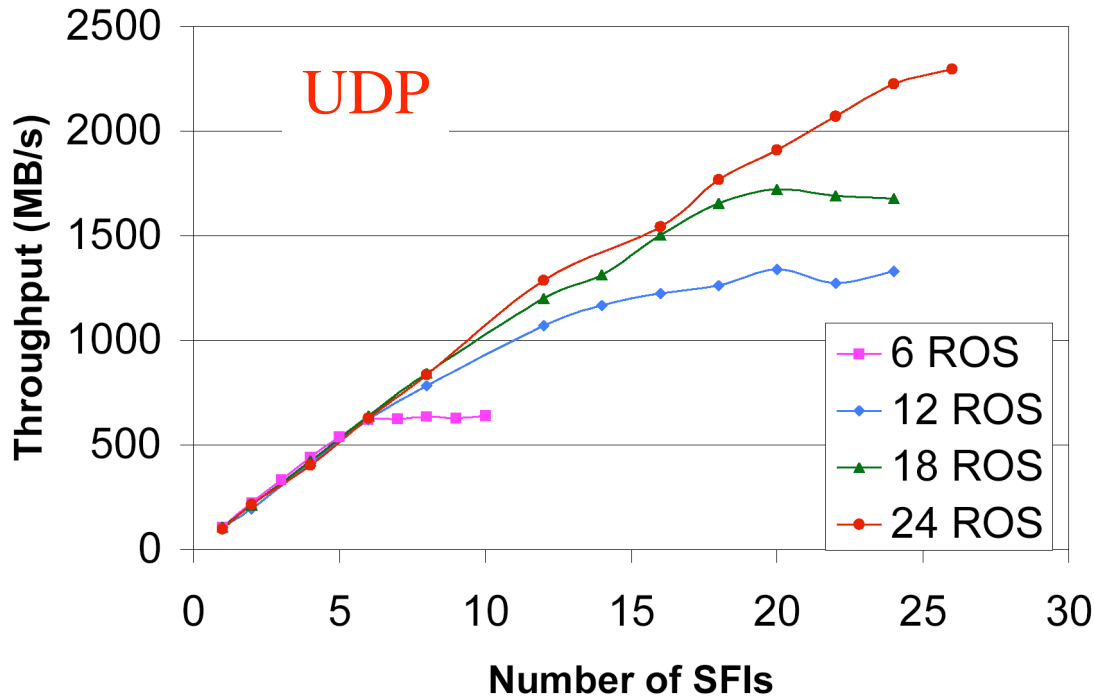
Modelling results obtained with at2sim, makes use of Ptolemy simulation environment, and using calibrated component models

Results for event building obtained with 6 ROSs

Emulated ROBINs, 12 ROLs per ROS, two switches: BATM T6 and FastIron800, raw Ethernet, flow control on, max. 20 outstanding requests per SFI



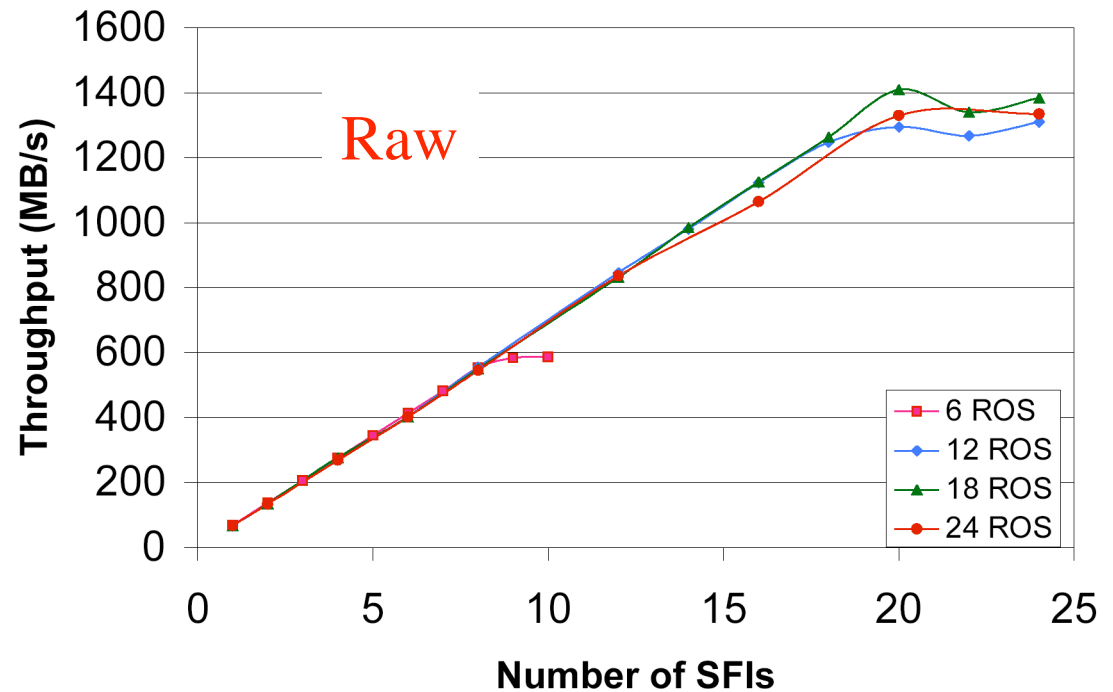
Modelling results obtained with at2sim, results with T6 switch not as expected



More results for event building obtained with 6 - 24 ROSs

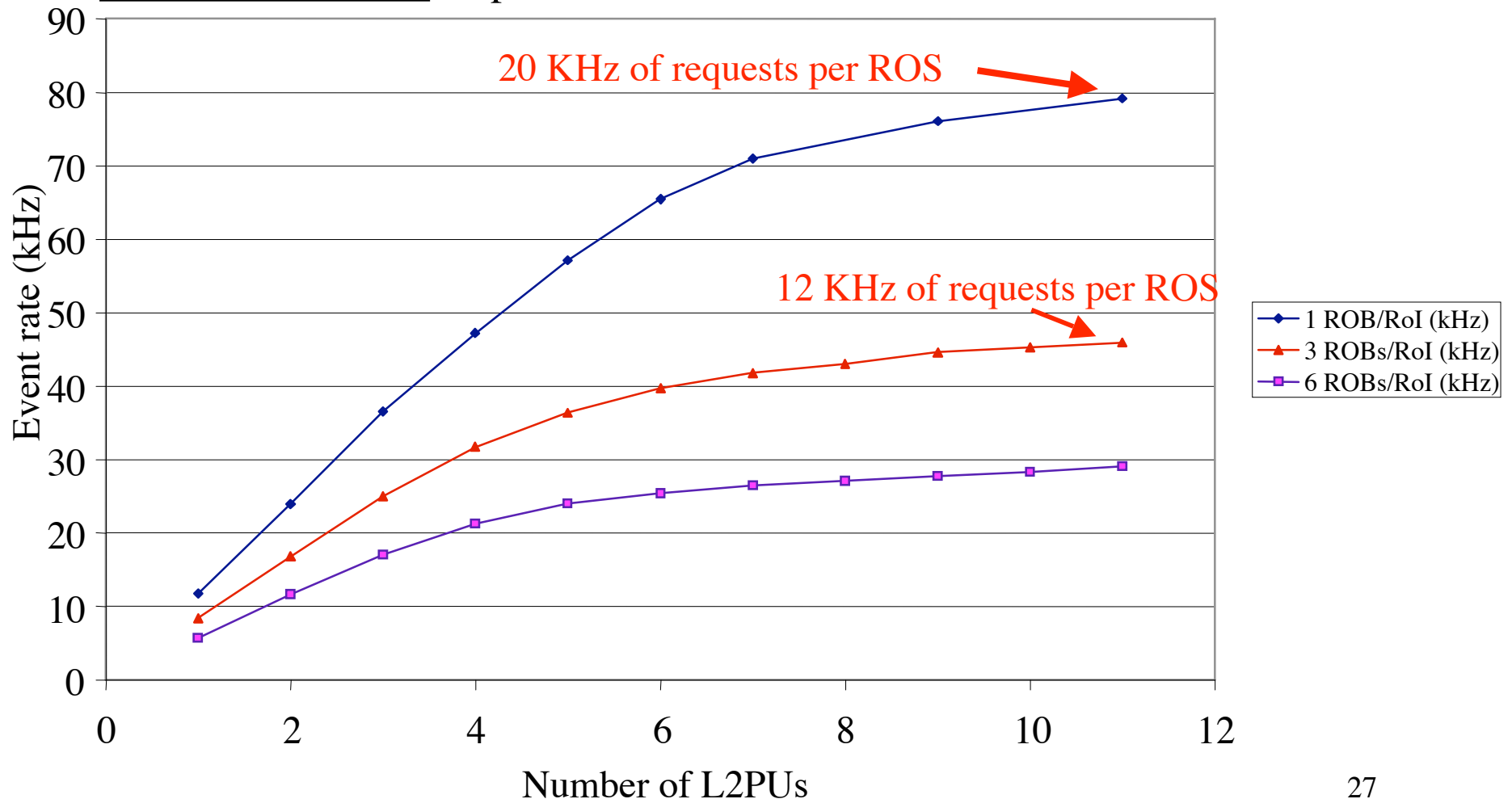
(with emulated ROBINs, 12 ROLs per ROS), FastIron 800 switch, flow control off, max. number of outstanding requests per SFI = 10

Expect saturation at about $N * 100$ MByte/s with N the number of ROSs

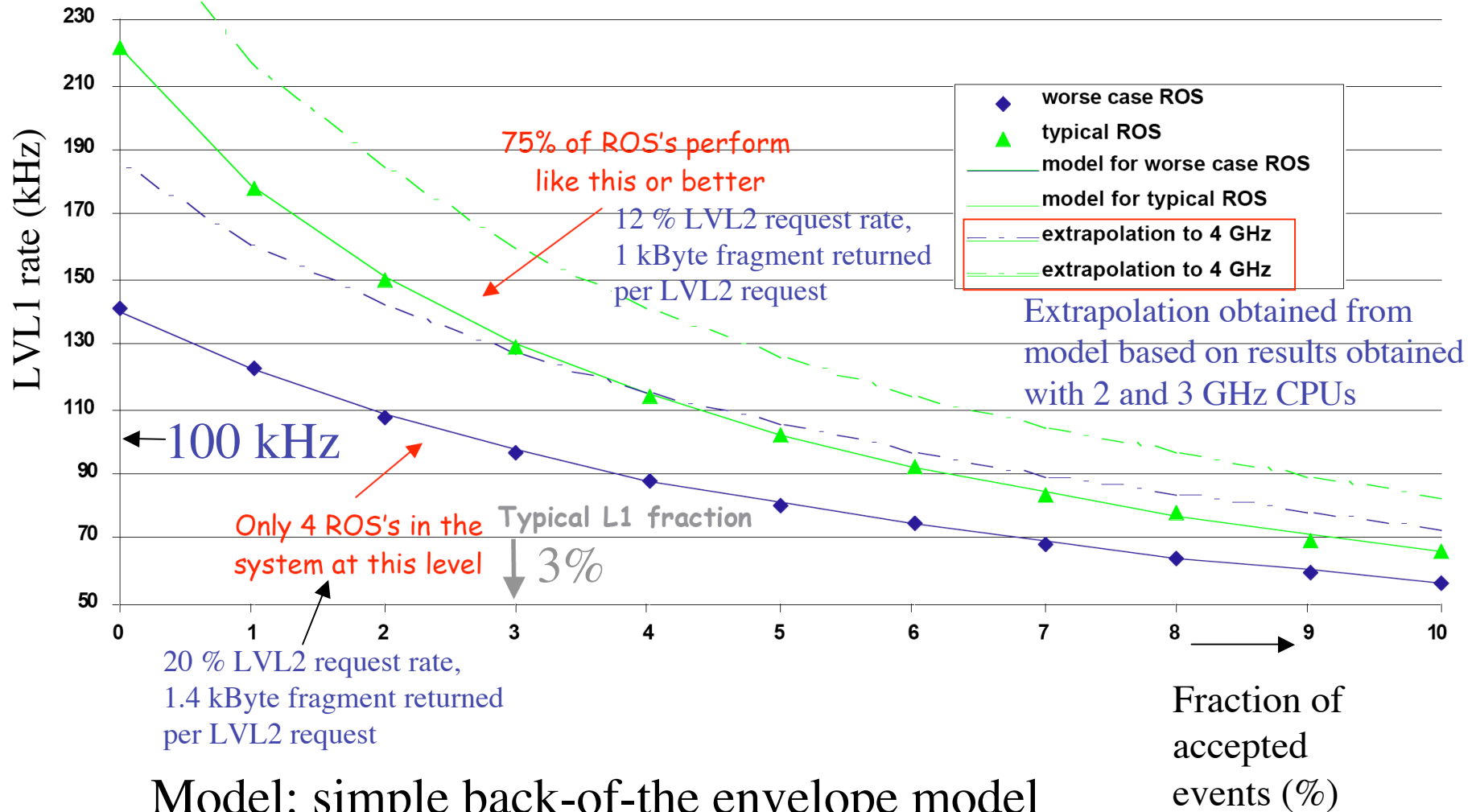


ROI request scalability

- 1-11 L2PUs (no algorithms) fetching data from 4 ROSs (12 inputs each)
- Different curves corresponds to different ROI sizes, 1.4 kByte per ROB
- 2.2 GHz machines, response of real ROBINS emulated



With 3 GHz PCs, three ROBIN emulators on PCI bus, 4 inputs per emulator (12 ROLs/ROS), 1 NIC/ROS (2 in final system)



Conclusions

Implementation:

- Standard rack-mounted PCs running Linux, software: multi-threaded C++
- Gigabit Ethernet networking
- Only dedicated hardware used for RoI Builder and ROBINS

The system design is complete, optimisation possible:

- of the I/O at the Read-Out System level
- of the deployment of the LVL2 and Event Builder networks

The functionality and performance of the architecture has been validated via:

- deployment of full systems:
 - On testbed prototypes
 - At the ATLAS H8 test beam (not reported in this contribution)
- modelling

The architecture allows for deferring the purchase of part of the system and upgrading its rate capability in a later stage

Further testbed and modelling studies under way to ensure the absence of potential problems