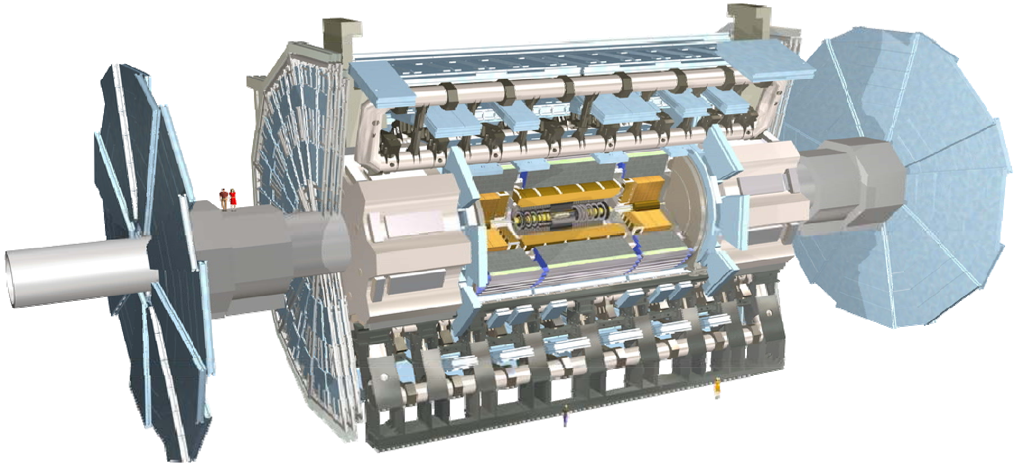
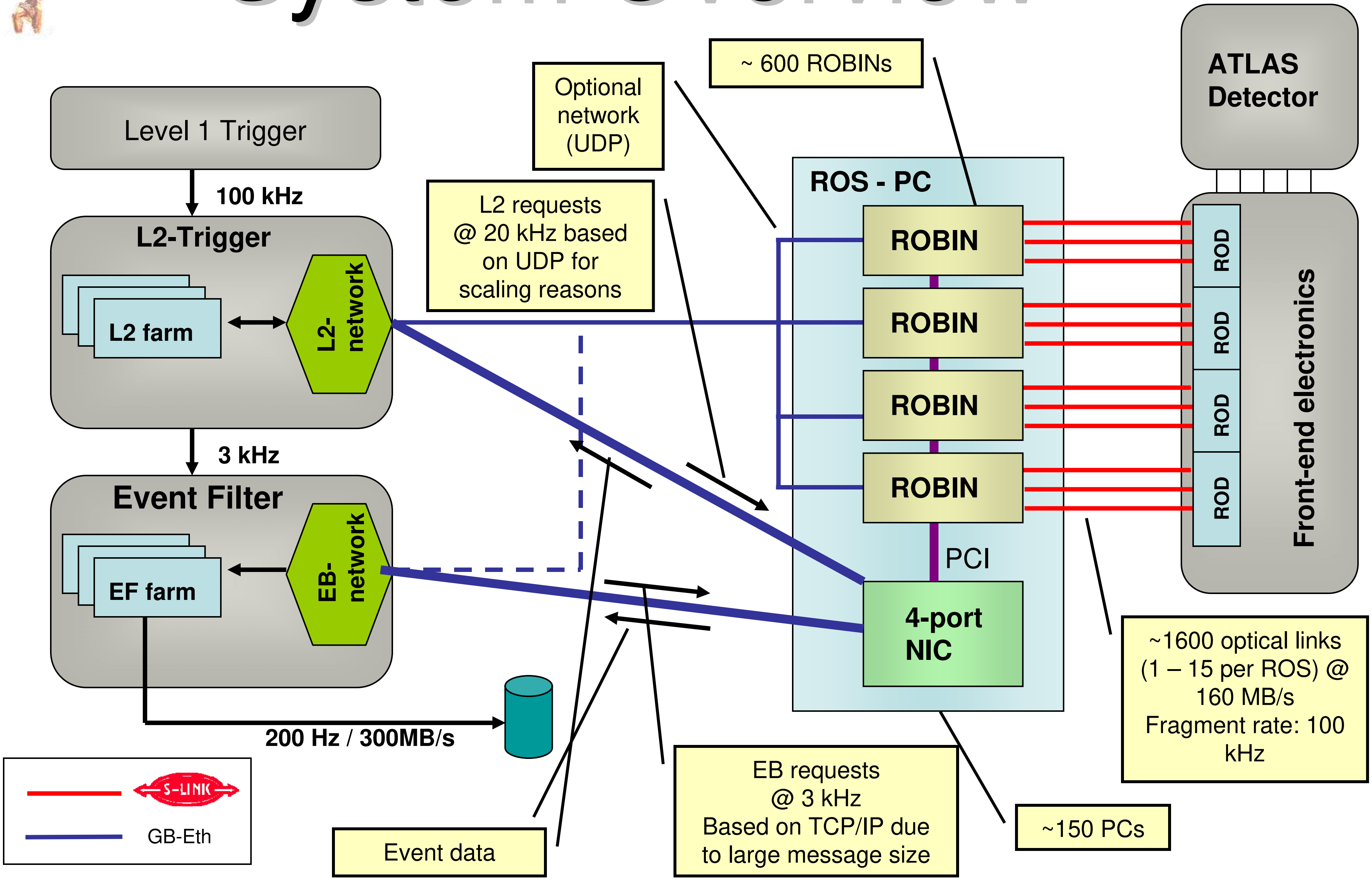
The ATLAS Read-Out System
Performance with first data and perspective for the future

A. Misiejuk[1], G. Crone[2], D. Della Volpe[3], B. Gorini[4], B. Green[1], **M. Joos**[4], G. Kieft[5], K. Kordas[8], A. Kugel[6], N. Schroer[6], P. Teixeira-Dias[1], L. Tremblet[4], J. Vermeulen[5], F. Wickens[7,] P. Werner[4]

[1]Royal Holloway University of London, [2]University College London, [3]Universita & INFN, Napoli, [4]CERN, [5]Nikhef, Amsterdam, [6]Ruprecht-Karls-Universitaet Heidelberg, [7]Rutherford Appleton Laboratory, [8]University Bern



System Overview



Level 1 Trigger

100 kHz

L2-Trigger

L2 farm

L2-network

3 kHz

Event Filter

EF farm

EB-network

200 Hz / 300MB/s

Optional network (UDP)

L2 requests @ 20 kHz based on UDP for scaling reasons

~ 600 ROBINs

ROS - PC

ROBIN

ROBIN

ROBIN

ROBIN

PCI

4-port NIC

ATLAS Detector

ROD

ROD

ROD

ROD

Front-end electronics

~1600 optical links (1 – 15 per ROS) @ 160 MB/s Fragment rate: 100 kHz

~150 PCs

EB requests @ 3 kHz Based on TCP/IP due to large message size

Event data

S-LINK

GB-Eth

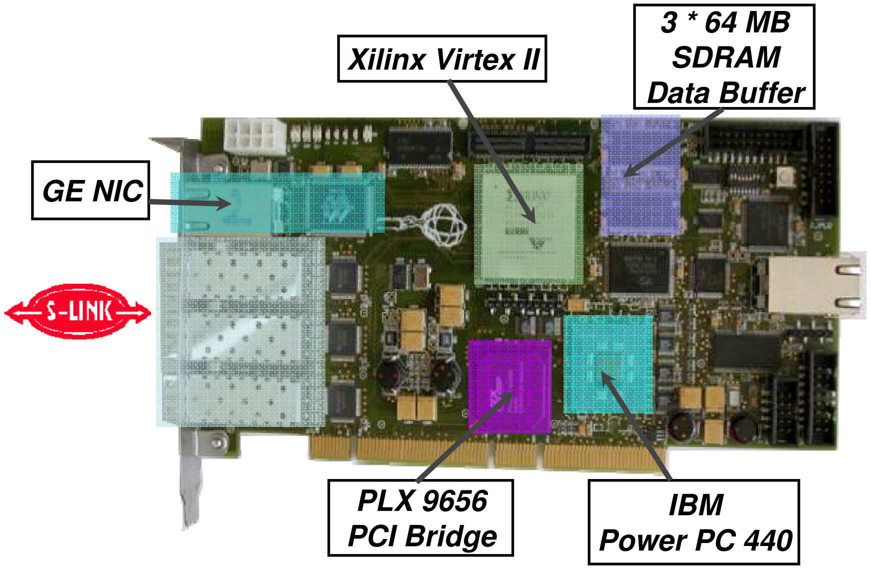
Building blocks

The ReadOut System (ROS) PC

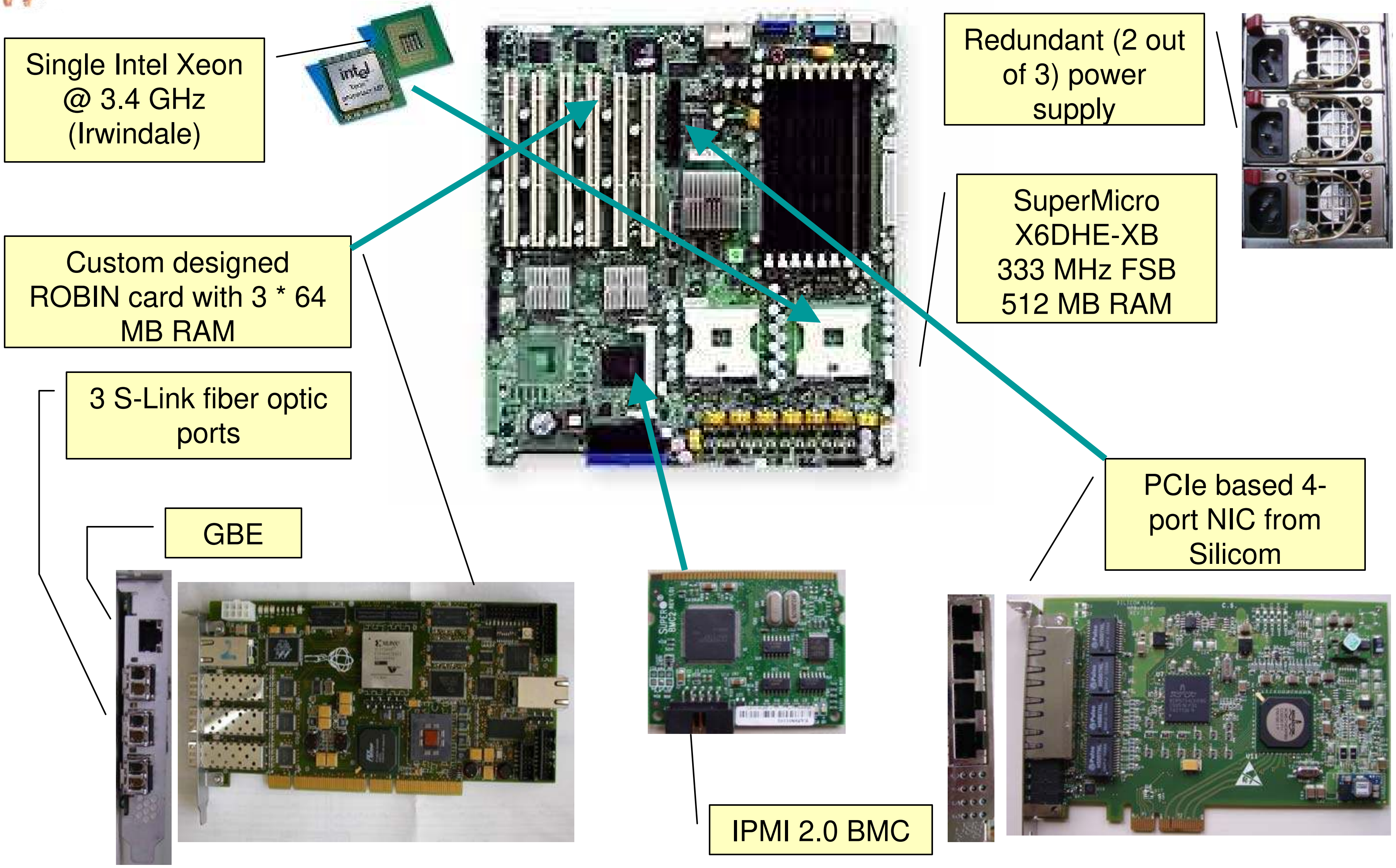
- Houses 1 to 5 ROBIN cards (typically 4 cards)
- Configures and controls the ROBINs
- Reads data from the ROBINs and provides it to the Second-Level Trigger and to the Event Builder
- Receives clear requests for event fragments and forwards them to the ROBINs
- Interfaces to the operational and physics monitoring systems

The ROBIN PCI card

- Receives event fragments from sub-detector specific front-end electronics (RODs) via 3 optical links
- Buffers events during the decision latency of the Second-Level Trigger and the time required for building of events accepted by L2
- The optical link is based on the S-LINK interface: 32 bit @ 40MHz = 160 MB/s



Hardware Components



Single Intel Xeon @ 3.4 GHz (Irwindale)

Custom designed ROBIN card with 3 * 64 MB RAM

3 S-Link fiber optic ports

GBE

Redundant (2 out of 3) power supply

SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM

PCIe based 4-port NIC from Silicom

IPMI 2.0 BMC

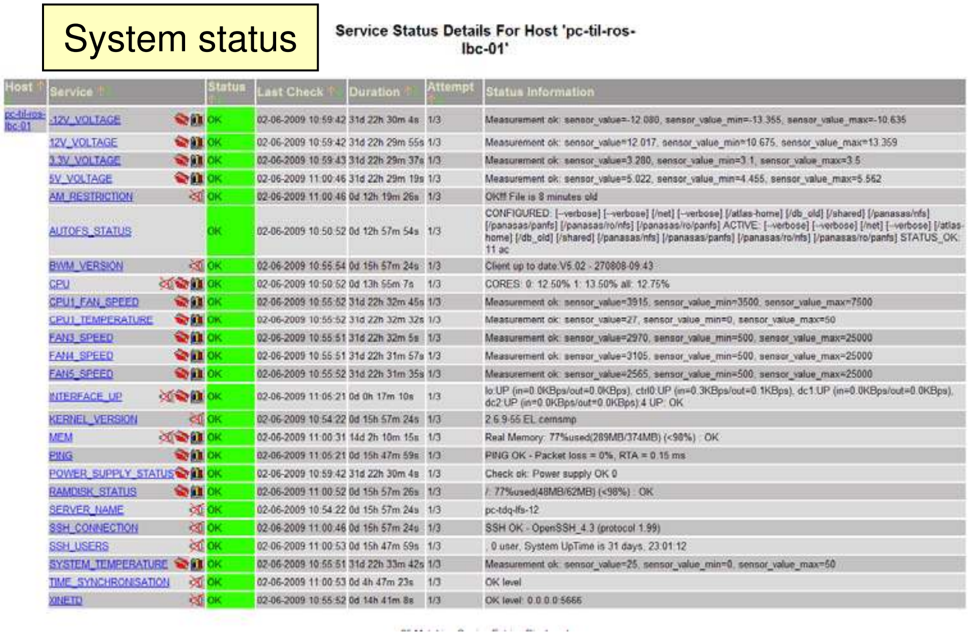
Hardware Components

15/07/2006

68 ROS PC (liquid Argon sub detector)

Operational Monitoring

System status



Service Status Details For Host 'pc-til-ros-lbc-01'

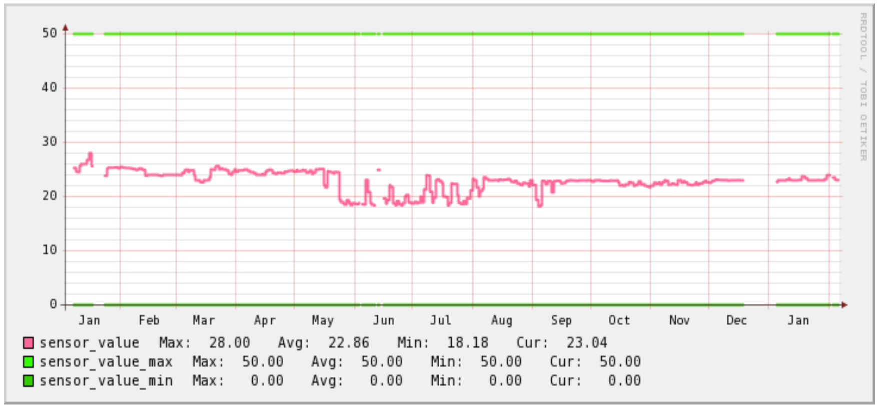
Host	Service	Status	Last Check	Duration	Attempt	Status Information
pc-til-ros-lbc-01	-12V_VOLTAGE	OK	02-06-2009 10:59:42	31d 22h 30m 4s	1/3	Measurement ok: sensor_value=-12.080, sensor_value_min=-13.355, sensor_value_max=-10.635
	12V_VOLTAGE	OK	02-06-2009 10:59:42	31d 22h 29m 55s	1/3	Measurement ok: sensor_value=12.017, sensor_value_min=10.675, sensor_value_max=13.359
	3.3V_VOLTAGE	OK	02-06-2009 10:59:43	31d 22h 29m 37s	1/3	Measurement ok: sensor_value=3.280, sensor_value_min=3.1, sensor_value_max=3.5
	5V_VOLTAGE	OK	02-06-2009 11:00:46	31d 22h 29m 19s	1/3	Measurement ok: sensor_value=5.022, sensor_value_min=4.455, sensor_value_max=5.562
	AM_RESTRICTION	OK	02-06-2009 11:00:46	0d 12h 19m 26s	1/3	OK!! File is 8 minutes old
	AUTOFS_STATUS	OK	02-06-2009 10:50:52	0d 12h 57m 54s	1/3	CONFIGURED: [--verbose] [--verbose] [/net] [--verbose] [/atlas-home] [/db_old] [/shared] [/panasas/nfs] [/panasas/panfs] [/panasas/ro/nfs] [/panasas/ro/panfs] ACTIVE: [--verbose] [--verbose] [/net] [--verbose] [/atlas-home] [/db_old] [/shared] [/panasas/nfs] [/panasas/panfs] [/panasas/ro/nfs] [/panasas/ro/panfs] STATUS_OK: 11 ac
	BWM_VERSION	OK	02-06-2009 10:55:54	0d 15h 57m 24s	1/3	Client up to date:V5.02 - 270808-09:43
	CPU	OK	02-06-2009 10:50:52	0d 13h 55m 7s	1/3	CORES: 0: 12.50% 1: 13.50% all: 12.75%
	CPU1_FAN_SPEED	OK	02-06-2009 10:55:52	31d 22h 32m 45s	1/3	Measurement ok: sensor_value=3915, sensor_value_min=3500, sensor_value_max=7500
	CPU1_TEMPERATURE	OK	02-06-2009 10:55:52	31d 22h 32m 32s	1/3	Measurement ok: sensor_value=27, sensor_value_min=0, sensor_value_max=50
	FAN3_SPEED	OK	02-06-2009 10:55:51	31d 22h 32m 5s	1/3	Measurement ok: sensor_value=2970, sensor_value_min=500, sensor_value_max=25000
	FAN4_SPEED	OK	02-06-2009 10:55:51	31d 22h 31m 57s	1/3	Measurement ok: sensor_value=3105, sensor_value_min=500, sensor_value_max=25000
	FAN5_SPEED	OK	02-06-2009 10:55:52	31d 22h 31m 35s	1/3	Measurement ok: sensor_value=2565, sensor_value_min=500, sensor_value_max=25000
	INTERFACE_UP	OK	02-06-2009 11:05:21	0d 0h 17m 10s	1/3	lo:UP (in=0.0KBps/out=0.0KBps), ctrl0:UP (in=0.3KBps/out=0.1KBps), dc1:UP (in=0.0KBps/out=0.0KBps), dc2:UP (in=0.0KBps/out=0.0KBps):4 UP: OK
	KERNEL_VERSION	OK	02-06-2009 10:54:22	0d 15h 57m 24s	1/3	2.6.9-55.EL.cernsmp
	MEM	OK	02-06-2009 11:00:31	14d 2h 10m 15s	1/3	Real Memory: 77%used(289MB/374MB) (<98%) : OK
	PING	OK	02-06-2009 11:05:21	0d 15h 47m 59s	1/3	PING OK - Packet loss = 0%, RTA = 0.15 ms
	POWER_SUPPLY_STATUS	OK	02-06-2009 10:59:42	31d 22h 30m 4s	1/3	Check ok: Power supply OK 0
	RAMDISK_STATUS	OK	02-06-2009 11:00:52	0d 15h 57m 26s	1/3	/: 77%used(48MB/62MB) (<98%) : OK
	SERVER_NAME	OK	02-06-2009 10:54:22	0d 15h 57m 24s	1/3	pc-tdq-lfs-12
	SSH_CONNECTION	OK	02-06-2009 11:00:46	0d 15h 57m 24s	1/3	SSH OK - OpenSSH_4.3 (protocol 1.99)
	SSH_USERS	OK	02-06-2009 11:00:53	0d 15h 47m 59s	1/3	, 0 user, System UpTime is 31 days, 23:01:12
	SYSTEM_TEMPERATURE	OK	02-06-2009 10:55:51	31d 22h 33m 42s	1/3	Measurement ok: sensor_value=25, sensor_value_min=0, sensor_value_max=50
	TIME_SYNCHRONISATION	OK	02-06-2009 11:00:53	0d 4h 47m 23s	1/3	OK level
	XINETD	OK	02-06-2009 10:55:52	0d 14h 41m 8s	1/3	OK level: 0.0.0.0:5666

- Lowest level: IPMI & ssh
- Server level: Nagios
- User level: Web browser

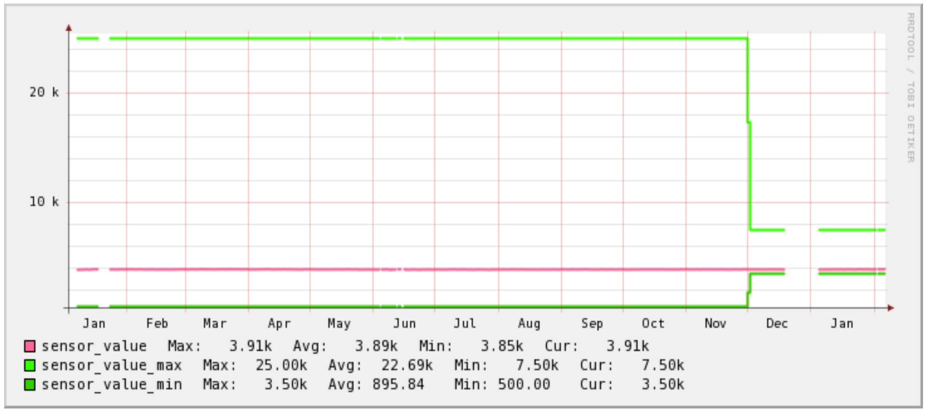
Features:

- History charts
- Automatic E-mail notification in case of problems

System temperature



Fan speed



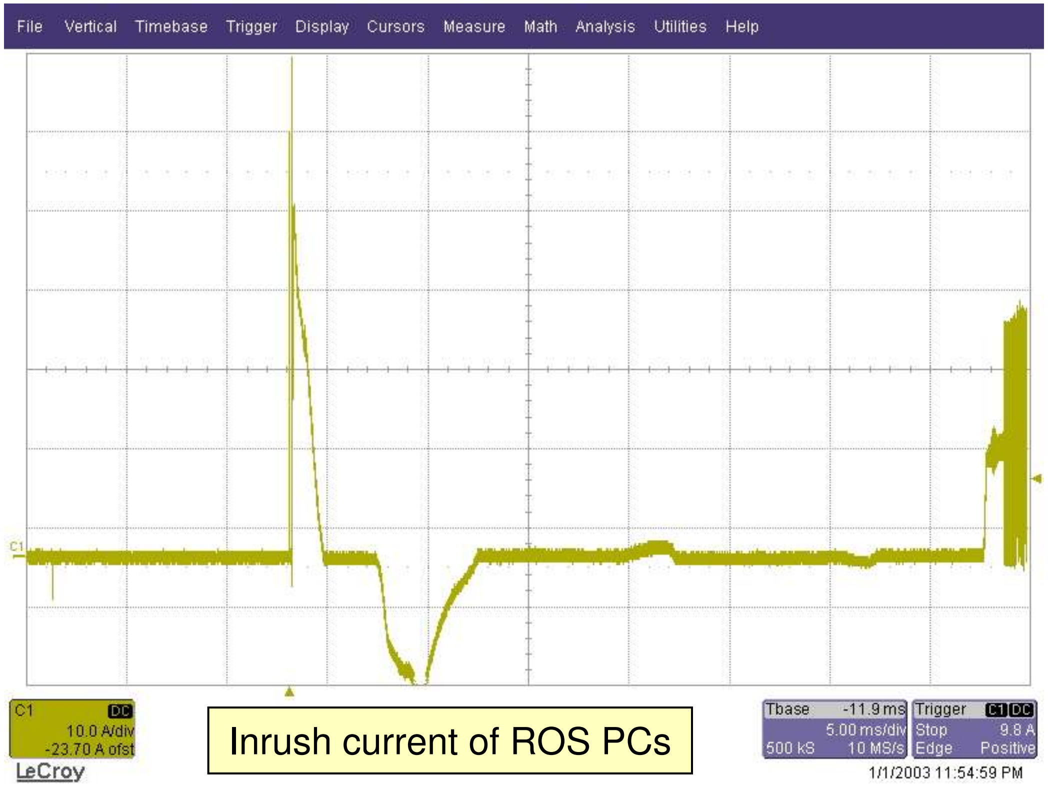
Hardware reliability

Type of component	Number of installed units	Number of broken units	Failures per year [%]	Failures in 2008
PC Motherboard	150	3	0.77	1
CPU	150	1	0.26	0
Memory DIMM	300	3	0.39	1
Power Supply module	450	4	0.34	1
IPMI BMC	150	4	1.03	0
CPU ventilator	150	2	0.51	0
chassis ventilator	450	1	0.09	0
4-port NIC	150	1	0.26	1
ROBIN cards - broken	614	23	1.45	1
ROBIN cards – intermittent errors (Firmware issues)	614	36	2.26	15

Average age of the hardware: 2.6 years

System integration issues

- The individual PSUs of the PCs generate a significant inrush current peak when power is restored after a power cut
- This may cause breakers in the rack to trip
- First solution: Power staggering barrettes
    - But recent questions on long-term reliability
- So now plan second solution
    - Thermistors



Inrush current of ROS PCs

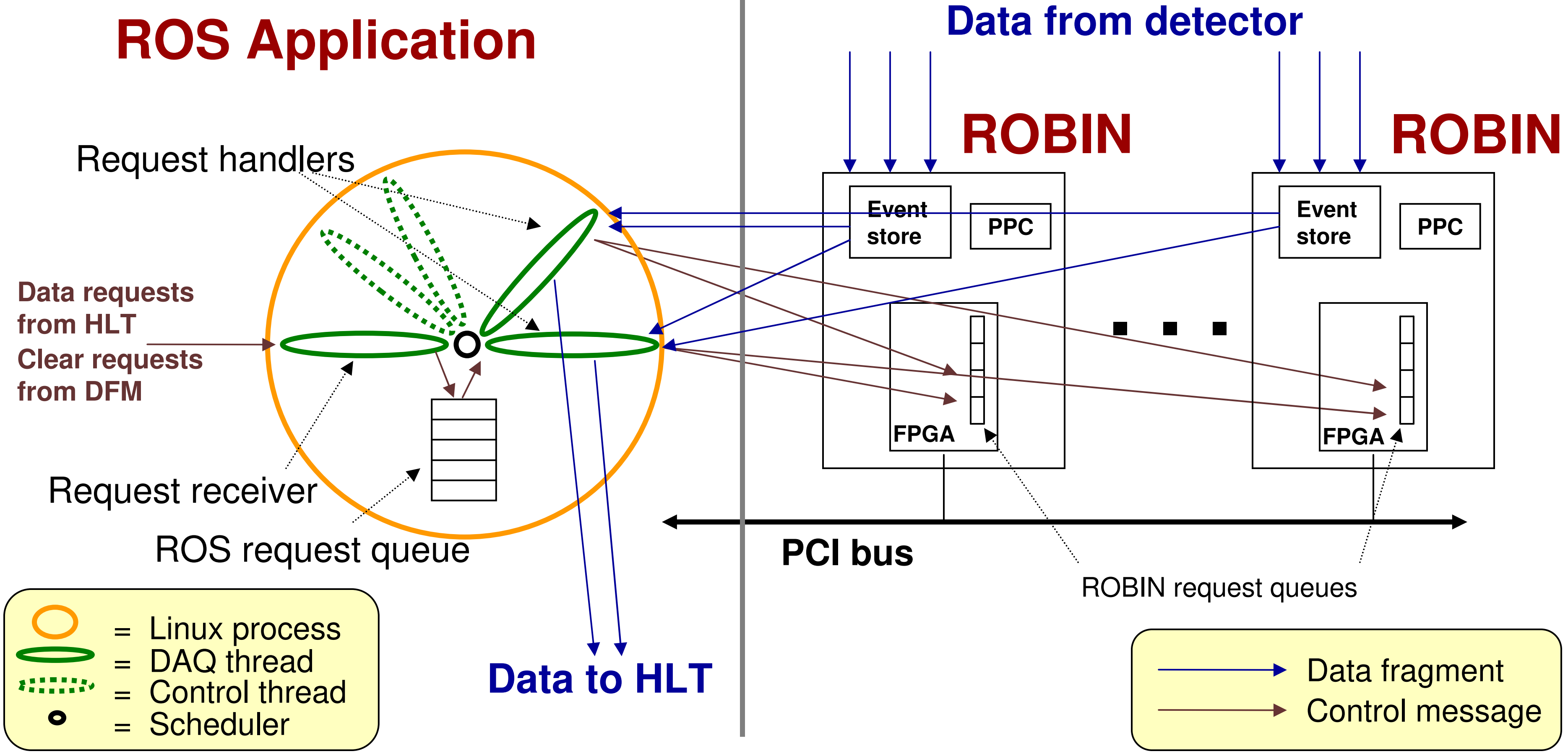
Power staggering barrette

ROS software Architecture

The application retrieves data fragments from the ROBINs, combines them in a unique fragment and sends it to L2/EB

Multi-threaded C++ program running under Linux (SLC4)

PowerPC processor in ROBIN runs C program booted from FLASH memory



ROS Application

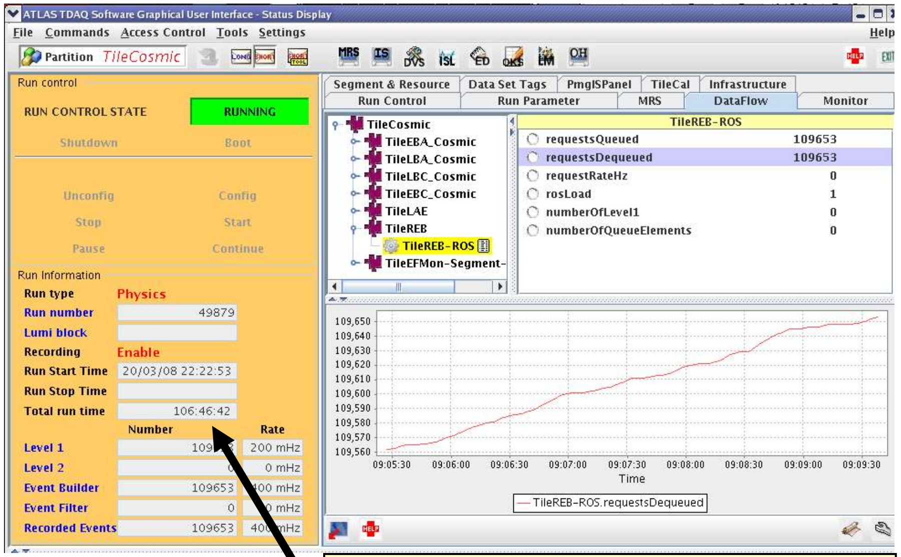
Data from detector

= Linux process
= DAQ thread
= Control thread
= Scheduler

Data fragment
Control message

System performance in 2008

- The ROS PCs were powered almost permanently
- Most of the time they were used for data taking with a trigger on cosmic events
- ROS data output statistics for August to October:
  - ~900 TB of data
- All detectors were commissioned at full S-Link (ROD-ROS) speed
- During selected periods high-rate tests were performed with pre-loaded data (ROBIN not involved)
  - 136 ROS PCs delivered 5.3 GB/s to HLT (40 MB/s/ROS)



Partition TileCosmic

Run control

RUN CONTROL STATE: RUNNING

Run Information

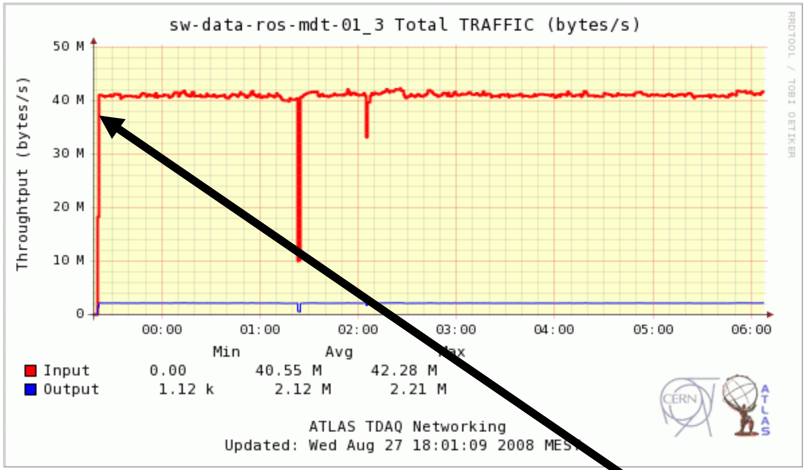
Run type	Physics	
Run number	49879	
Lumi block		
Recording	Enable	
Run Start Time	20/03/08 22:22:53	
Run Stop Time		
Total run time	106:46:42	
	Number	Rate
Level 1	109[illegible]	200 mHz
Level 2	[illegible]	0 mHz
Event Builder	109653	[illegible]00 mHz
Event Filter	0	[illegible] mHz
Recorded Events	109653	40[illegible]mHz

TileCosmic
- TileEBA_Cosmic
- TileLBA_Cosmic
- TileLBC_Cosmic
- TileEBC_Cosmic
- TileLAE
- TileREB
  - TileREB-ROS
- TileEFMon-Segment-

TileREB-ROS

○ requestsQueued	109653
○ requestsDequeued	109653
○ requestRateHz	0
○ rosLoad	1
○ numberOfLevel1	0
○ numberOfQueueElements	0

A continuous run of more than 4 days



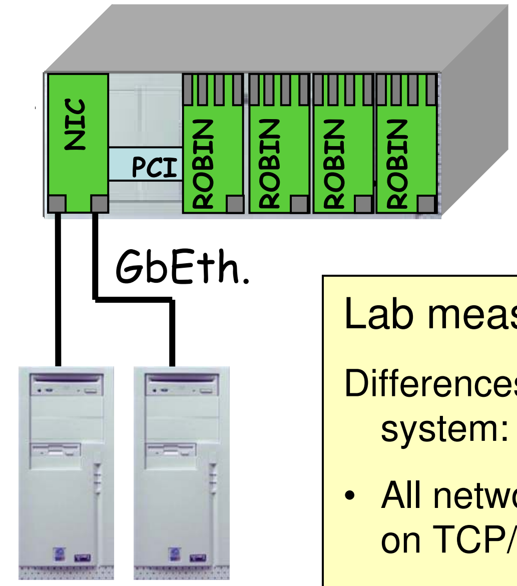
	Min	Avg	Max
🟥 Input	0.00	40.55 M	42.28 M
🟦 Output	1.12 k	2.12 M	2.21 M

ATLAS TDAQ Networking
Updated: Wed Aug 27 18:01:09 2008 MES[illegible]

Output of ROS to HLT during high rate test; 40 MB/s

Performance of the standard ROS

Test Setup

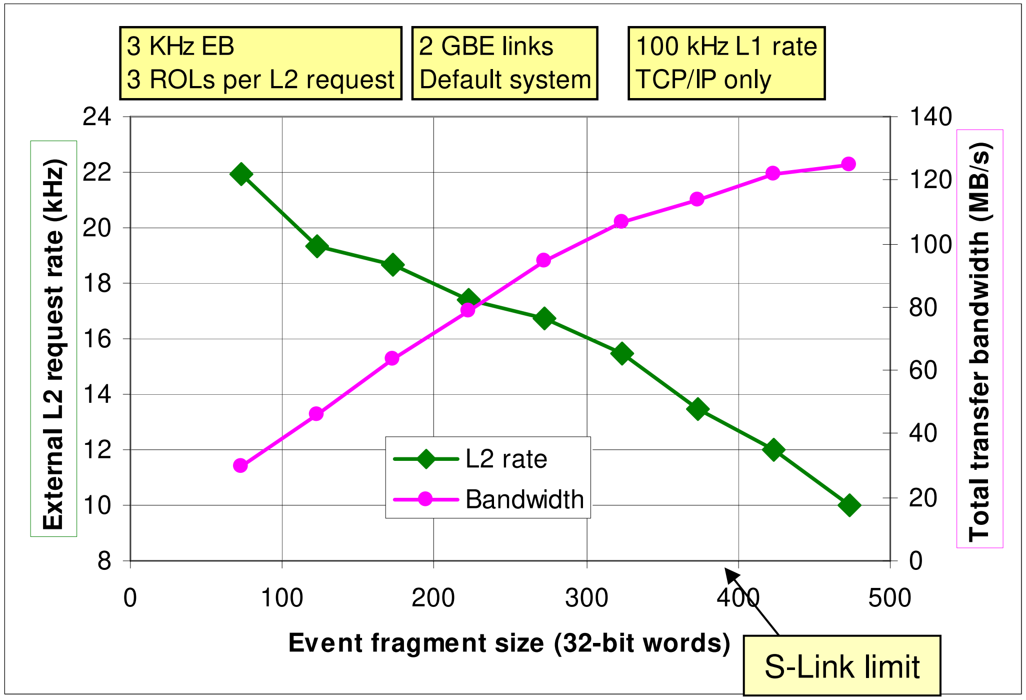


Requester PCs, emulate HLT nodes

Lab measurement

Differences wrt deployed system:

- All network traffic based on TCP/IP
- ROBINs generate data internally



3 KHz EB
3 ROLs per L2 request

2 GBE links
Default system

100 kHz L1 rate
TCP/IP only

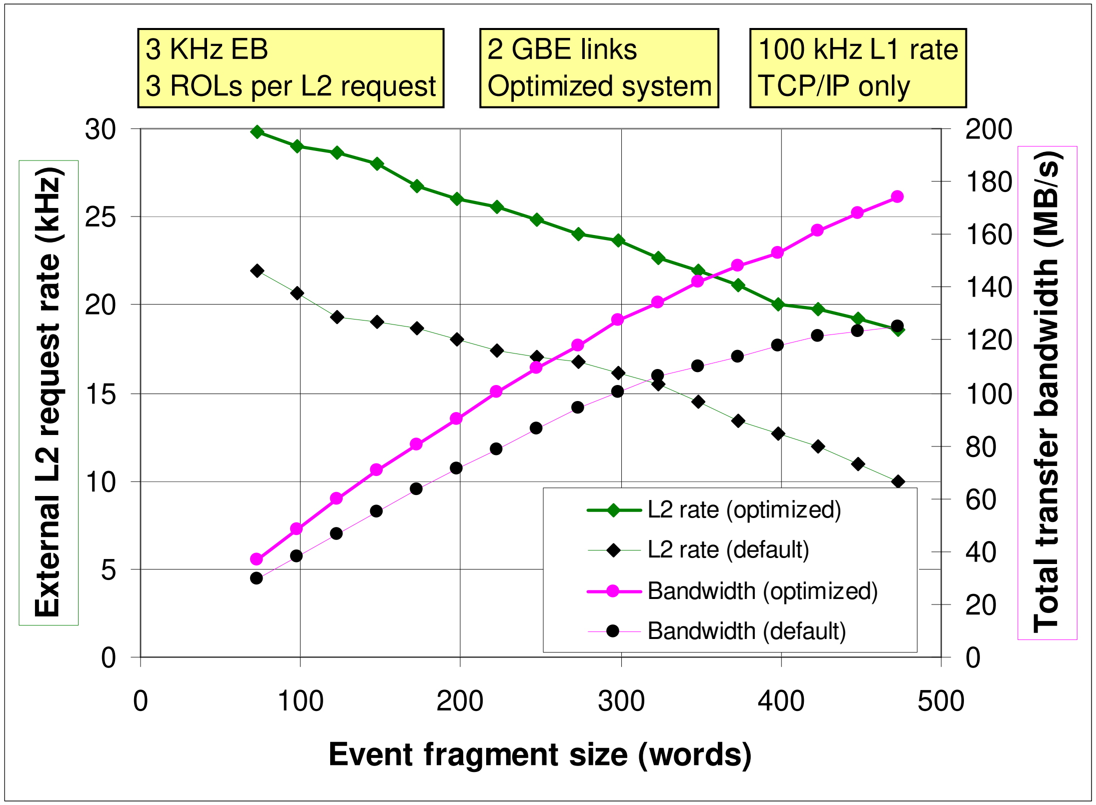
S-Link limit

- Canonical fragment size in ATLAS: ~ 1 kByte (256 words)
- Meets the original requirements (~20 kHz L2 rate) for small fragment sizes
- Too slow for large fragments
- Bottleneck seems to be the ROS CPU (ROBINs & NICs can sustain much higher rates)

Performance of the optimized ROS

The original configuration of the OS & drivers turned out to be inefficient.
Finally we obtained the best performance by:

- Turning hyperthreading off
- Using a uni-processor kernel
- Tuning the interrupt coalescence of the network driver
- Changing the SELinux configuration



- Performance is now OK for all fragment sizes
- However would like more headroom

Higher ROS performance - motivation

- ATLAS (upgrade) phases
  - Phase 0 (until 2013, luminosity: up to $1*10^{34}$ cm^{-2} s^{-1})
    - Need more ROS performance to:
      - have headroom for ROS PCs with high L2 request rates
      - compensate for higher rates due to modified thresholds of the L2 trigger
      - allow for additional bandwidth-demanding types of triggers. E.g.:
        - » Inner detector full scan for b-physics
        - » Calorimeter full scan for missing E_T
  - Phase 1 (2013 – 2017, luminosity: up to $3*10^{34}$ cm^{-2} s^{-1})
    - Higher data rates due to increased luminosity
    - Still use (current) ROS PCs & ROBINs
    - Requires more network bandwidth (switches, ROBINs & ROS)
  - Phase 2 (from 2018, luminosity: up to $10*10^{34}$ cm^{-2} s^{-1})
    - Much higher data rates
    - Replace ROS system

Higher ROS performance - options

(for phase 0 & 1)

- The main bottleneck of the ROS is the network interface to the HLT
  - Only 2 GBE links per ROS
  - Network protocol (mix of UDP and TCP/IP) handled by the CPU of the ROS PC
- 3 approaches to solve the network limitation

Install smart NICs (to offload CPU from the TCP/IP protocol)

Replace the motherboard, CPU and memory of the ROS PCs with faster hardware and connect additional GBE lines from the ROS PCs to the HLT network

Connect the ROBINs directly to the HLT network

Chelsio S320E 2-port smart NIC

SuperMicro X7DB8-X MB with 2 * 2.66 GHz quad core Xeon and RAM @ 667 MHz

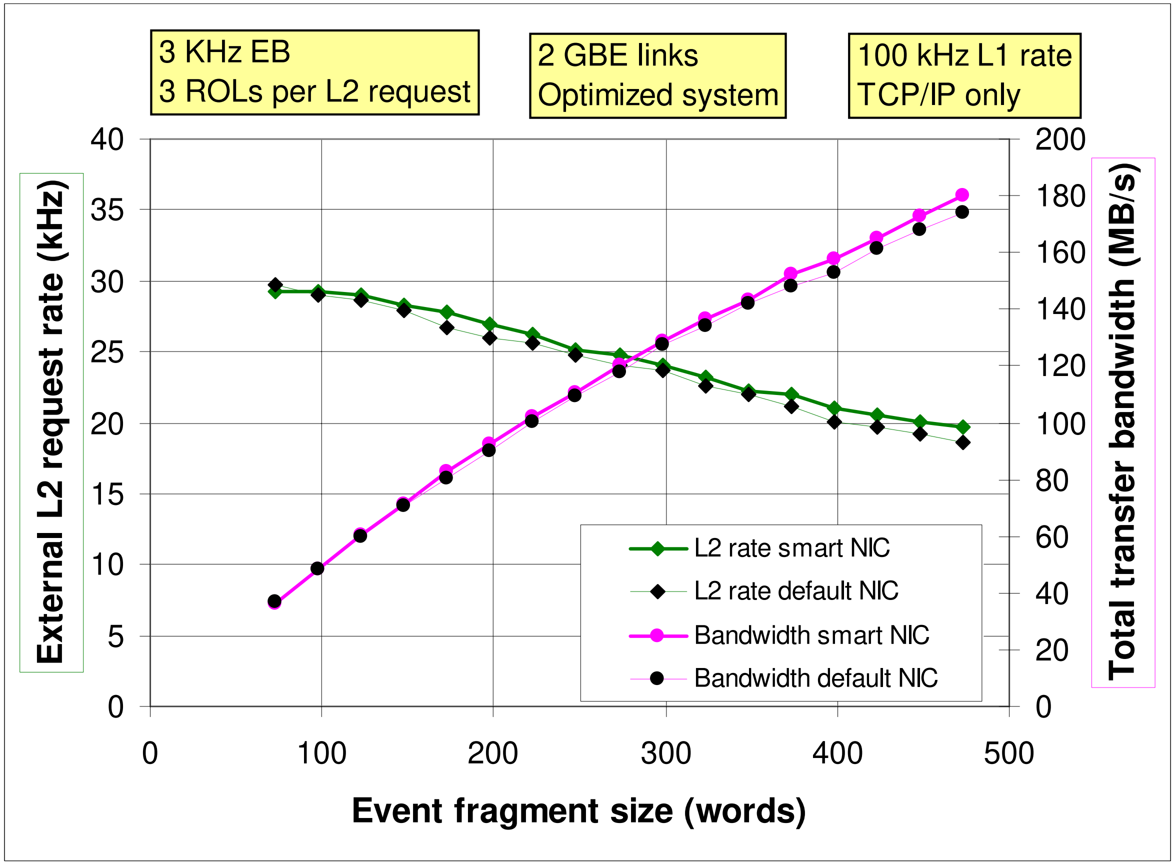
This may be one of the last MBs with >3 64bit PCI slots -> Development of PCIe based ROBIN started

GBE port of ROBIN (UDP only)

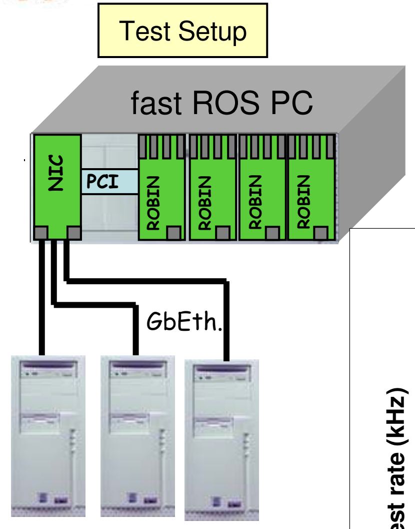
Impact of the Smart NIC

Laboratory measurement
- TCP/IP only

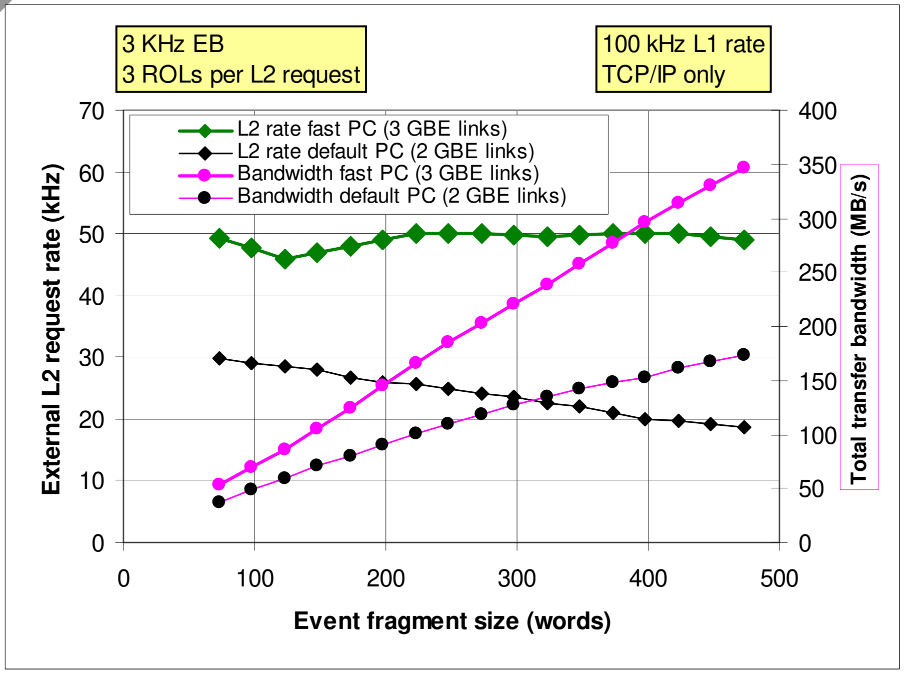
- No significant gain in performance
- Driver for smart NIC not yet tuned for best performance (waiting for optimized driver from manufacturer)
- More work required but overall potential seems to be low



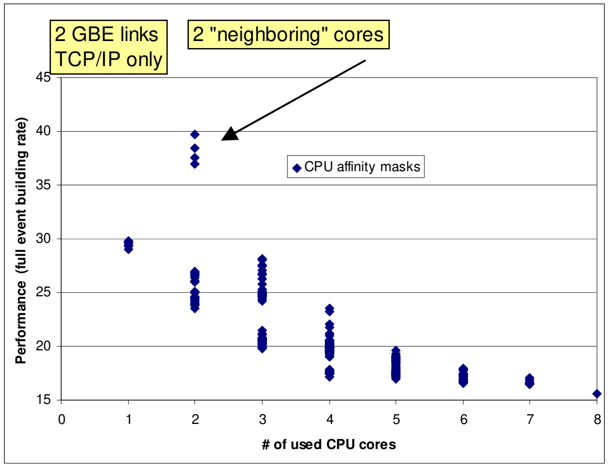
Impact of faster MB, CPU & RAM



Requester PCs, emulate HLT nodes


3 KHz EB
3 ROLs per L2 request

100 kHz L1 rate
TCP/IP only


2 GBE links
TCP/IP only

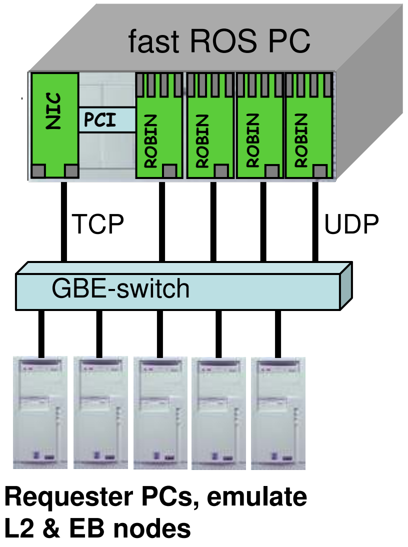
2 "neighboring" cores

Best performance with 2 (of 8) cores
-> effect of (extensive) use of mutexes?

- L2 request rate (almost) fragment size independent
- L2 request rate increases by 50% to 150%
- (expensive) ROBIN cards can be reused

Impact of Read-Out via PC & ROBIN cards

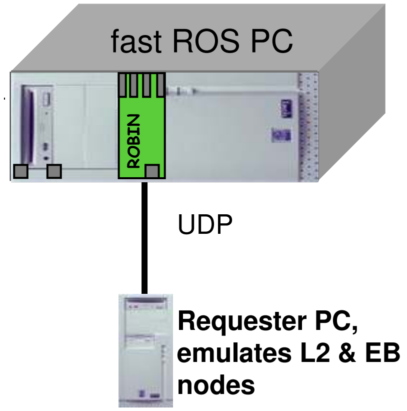
Test Setup



Full system test with (small HLT) farm
Only preliminary results so far

- Functionality OK
- current test system limits performance

Test Setup 2



Simplified set-up for tests at the ROBIN level

- This ROS configuration has the potential to deliver more performance than the ROS with the faster motherboard & CPU
- Further optimization of the system (software) required

Summary and Conclusions

- Since its installation in 2006/2007 the ROS system has worked very reliably
- The ROS in its current configuration meets the requirements that were specified in the ATLAS Technical Design Report
- Several alternatives exist for the further improvement of the performance
    - More detailed tests have to be carried out in the deployed system to better understand the relative advantages and disadvantages of these alternatives
- The development of a PCIe based ROBIN has been started
    - Because motherboards with at least 4 64-bit PCI slots become difficult to find
    - Faster PPC CPU will also improve ROBIN performance
- Based on today’s understanding of the ATLAS TDAQ (HLT rejection factor and algorithms) as well as the planned upgrades of ATLAS and LHC the current ROS architecture fulfills the requirements of phase 0 & 1 of ATLAS