TriDAS LHCC-Report
CMS Data Acquisition

S. Cittolin CERN, 19.1.99

DAQ overview
DAQ design status
Towards TDR and plan of work
Prototypes and milestones
Cost book, organization, calendar

The pdf file of this talk is available at:

http://cmsdoc.cern.ch/ftp/distribution/tridas/3.Misc/lhccdq0199.pdf

See also TirDAS home page at:

http://cmsdoc.cern.ch/ftp/afscms/TRIDAS/html/tridas.html

DAQ overview

Detector channels, data and trigger rates

Trigger system and DAQ structures

DAQ parameters and baseline design

CMS trigger levels

Trigger and data acquisition trends

Detector channels, data and trigger rates

Detector	Channels	Occupancy(%)	Event size(kB)
Pixel	80000000	.01	30
InnerTracker	16000000	3.	700
Preshower	512000	10.	50
Calorimeters	125000	5.	200
Muons	1000000	.1	10
Trigger			10

Trigger	Et Cufoff	Indiv. Rate	Cumul. Rate	Increm. Rate
Type	(GeV)	(kHz)	(kHz)	(kHz)
Sum Et	400	0.48	0.48	0.48
Miss Et	80	1.29	1.7	1.22
Single e	25	6.84	8.34	6.64
Double e	12	1.45	9.52	1.18
Single jet	100	2.06	10.7	1.16
Double jet	60	2.17	11.6	0.93
Triple jet	30	3.16	13.3	1.7
Quad jet	20	2.96	14.3	0.59
Jet + e	50 & 12	1.35	**14.9**	0.59
Single μ	20	7.8	7.8	7.8
Double μ	4	1.6	9.2	1.4
μ + e	4 + 8	5.5	14.4	5.2
μ + jet	4 + 40	0.3	14.4	<0.1
μ + Miss Et	4 + 60	1.0	15.3	0.9
μ + Sum Et	4 + 250	0.2	**15.3**	<0.1

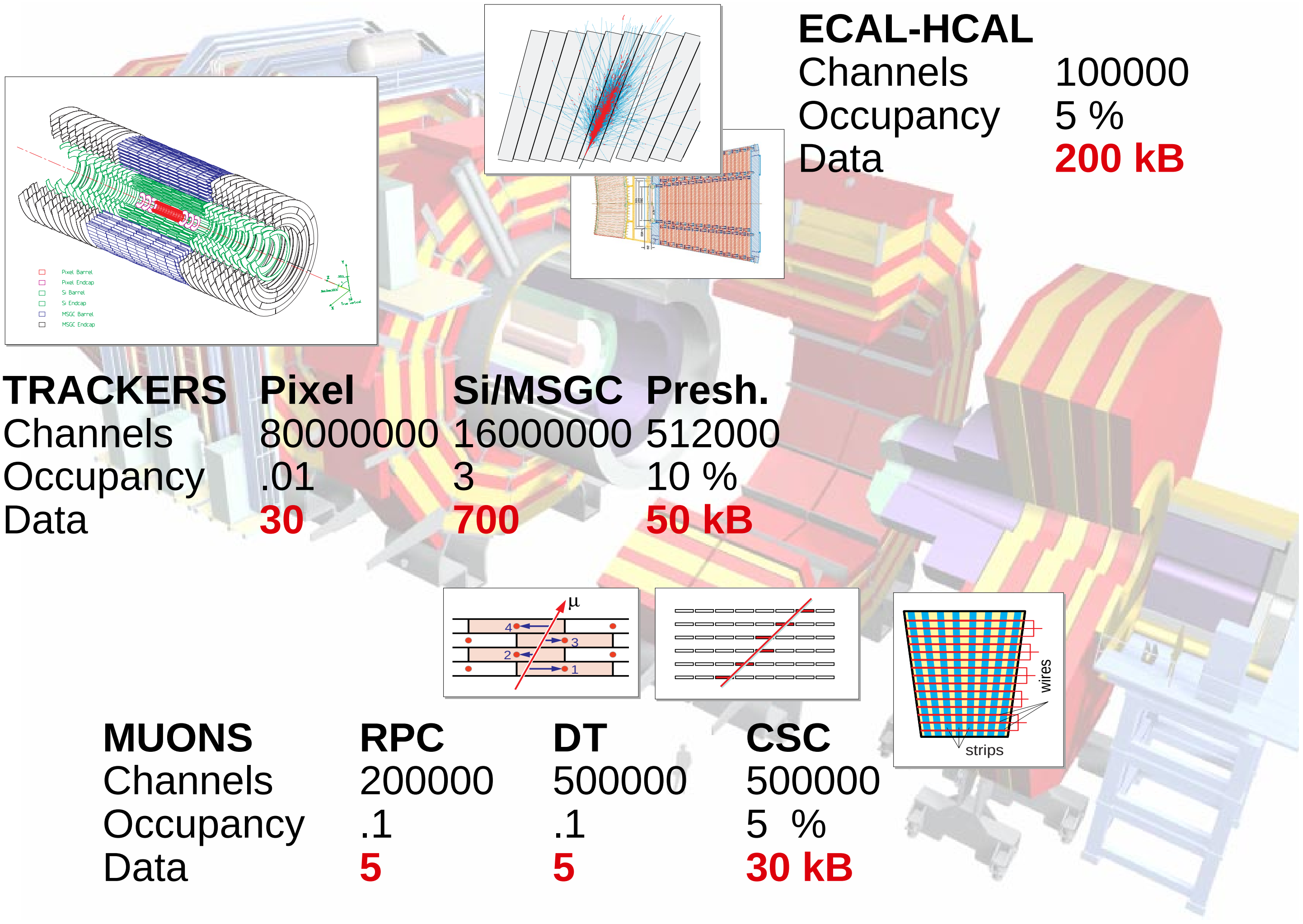
Trigger and Data Acquisition volume and rate requirements:

EVENT SIZE ~ **1 MByte**

PHYSICS TRIGGER RATE ≥ **30 kHz**

THE CMS TRIGGER AND DATA ACQUISITION SYSTEMS ARE DESIGNED TO OPERATE UP TO 100 kHz LEVEL-1 RATE WITH EVENT SIZE OF ABOUT 1 Mbyte

Detector channels, data and trigger rates



ECAL-HCAL	
Channels	100000
Occupancy	5 %
Data	200 kB

TRACKERS	Pixel	Si/MSGC	Presh.
Channels	80000000	16000000	512000
Occupancy	.01	3	10 %
Data	30	700	50 kB

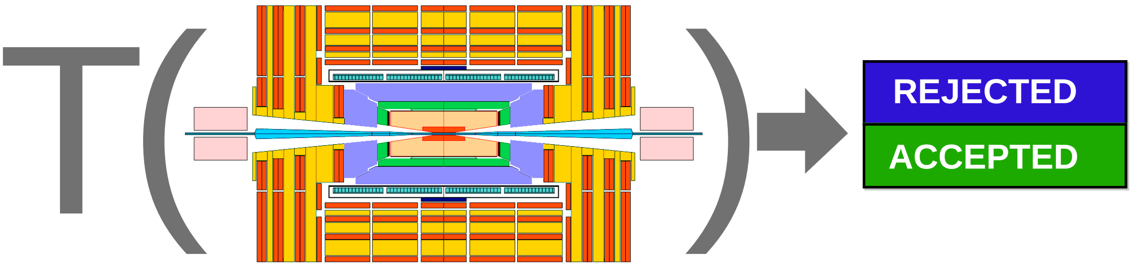
MUONS	RPC	DT	CSC
Channels	200000	500000	500000
Occupancy	.1	.1	5 %
Data	5	5	30 kB

Trigger and Data Acquisition volume and rate requirements:

EVENT SIZE	**~ 1 MByte**
PHYSICS TRIGGER RATE	**≥ 30 kHz**

Trigger system and DAQ structures

The trigger is a function of :



T( )

REJECTED

ACCEPTED

Event data & Apparatus
Physics channels & Parameters

Since the detector data are not all promptly available and the function is highly complex, T(...) is evaluated by successive approximations called :

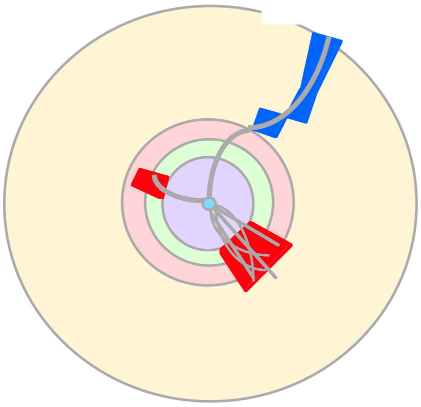
TRIGGER LEVELS

(possibly with zero dead time)

The trigger levels at LHC

Collision rate 10^{9} Hz

10^{-6} s

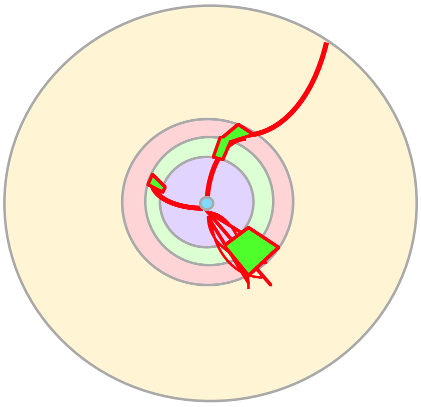


Particle identification (High p_T electron, muon, jets, missing E_T)

- Local pattern recognition and energy evaluation on prompt macro-granular information

Level-1 selected events 10^{5} Hz

10^{-3} s

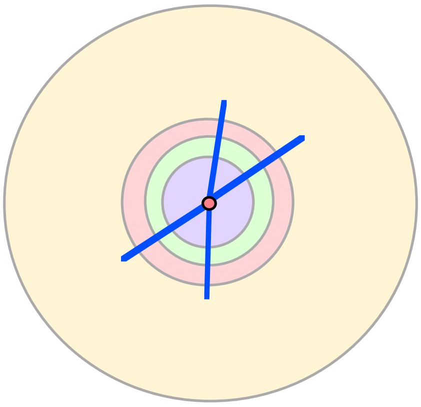


Clean particle signature (Z, W, ..)

- Finer granularity precise measurement
- Kinematics. effective mass cuts and event topology
- Track reconstruction and detector matching

Level-2 selected events 10^{3} Hz

10^{-0} s

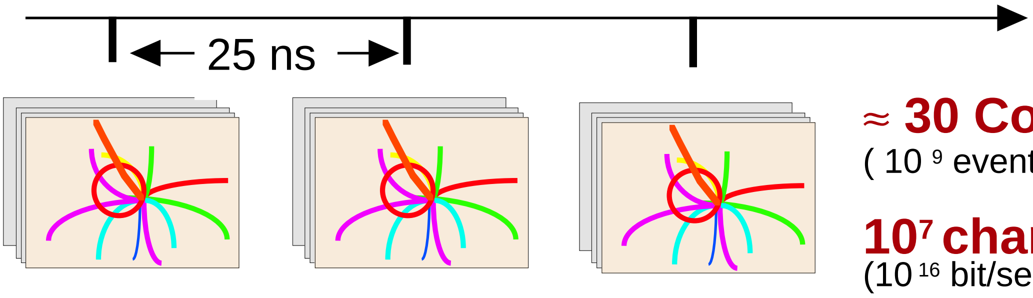


Physics process identification

- Event reconstruction and analysis

Level-3 events to tape 10..100 Hz

Trigger and readout structures at LHC

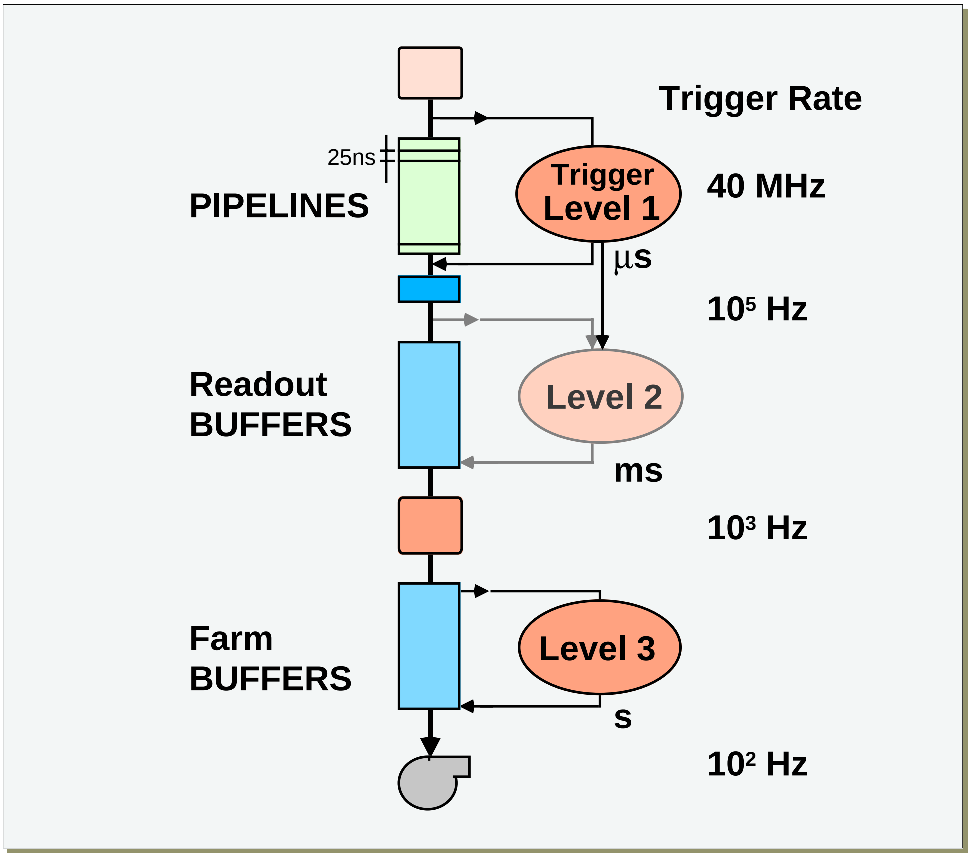


≈ **30 Collisions/25ns**
(10^9 event/sec)

10^7 channels
(10^{16} bit/sec)

Luminosity = 10^{34} cm^{-2} sec^{-1}

Multilevel trigger and readout systems



Trigger Rate

25ns

PIPELINES

Trigger Level 1

40 MHz

µs

10^5 Hz

Readout BUFFERS

Level 2

ms

10^3 Hz

Farm BUFFERS

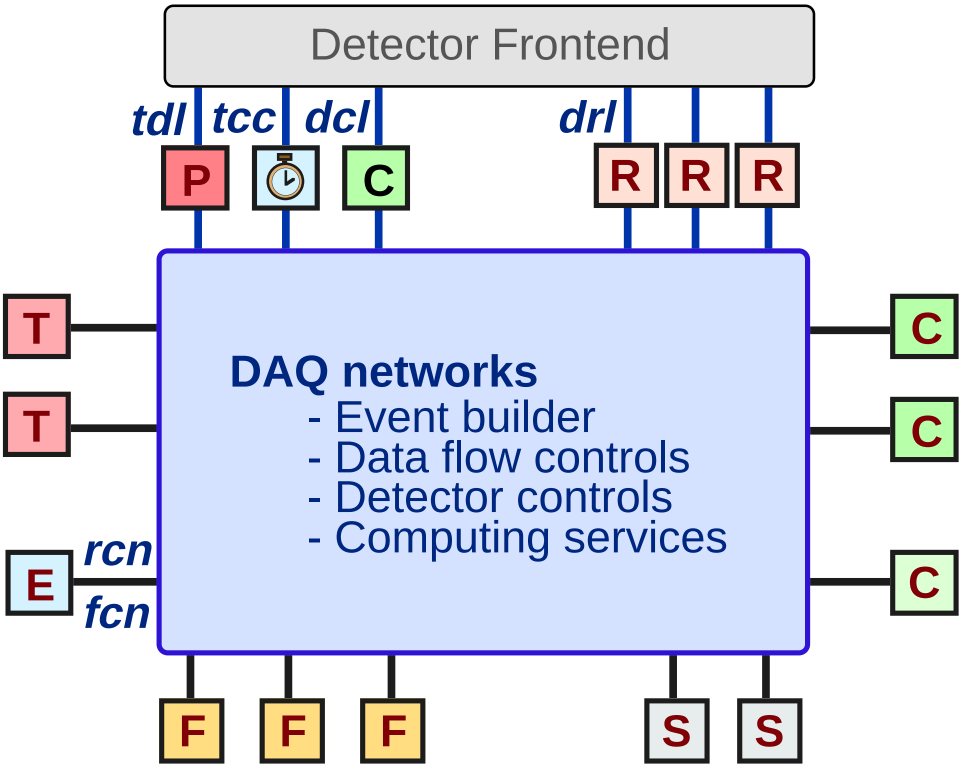
Level 3

s

10^2 Hz

10^7 channels x 25 ns data sampling

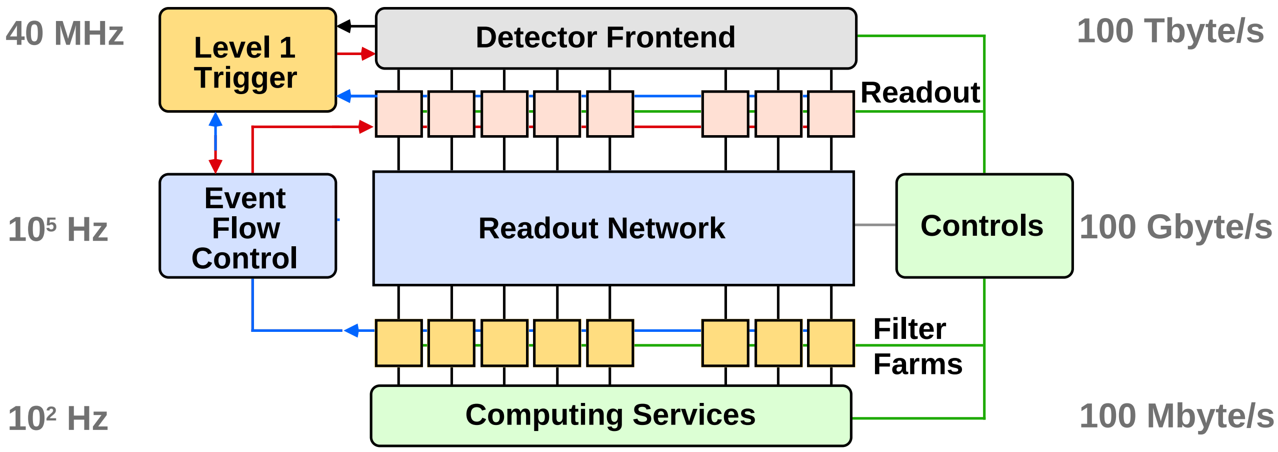
Trigger and data acquisition computing&communication systems



COMPUTING SYSTEMS	
P	:Primitive generators
T	:Trigger processors
E	:Event Flow Controls
C	:Detector Controllers
R	:Readout data formatters
F	:Event Filters
S	:Computing Services

COMMUNICATION NETWORKS	
tdl	:Trigger data links
tcc	:Timing and fast signals
dcl	:Detector control links
drl	:Detector readout links
rcn	:Readout control network
fcn	:Filter control network
csn	:Computing services network

DAQ parameters and baseline design

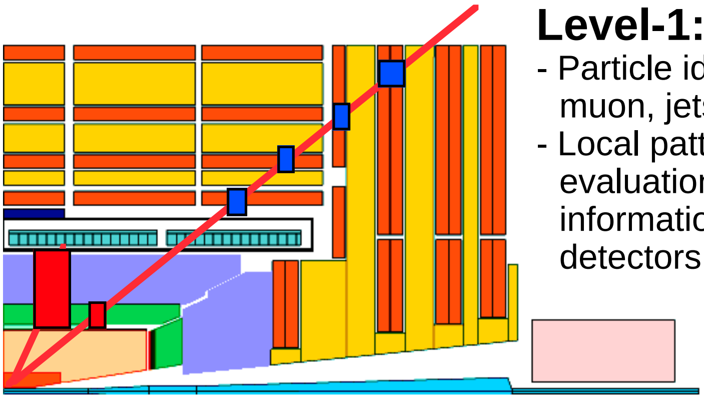


Collision rate	40 MHz
Level-1 Maximum trigger rate	100 kHz(*)
Average event size	≈ 1 Mbyte
Event Flow Control	≈ 10^6 Mssg/s
No. of In-Out units (200-5000 byte/event)	1000
Readout network (512-512 switch) bandwidth	≈ 500 Gbit/s
Event filter computing power	≈ 5 10^6 MIPS
Data production	≈ Tbyte/day
No. of readout crates	≈ 250
No. of electronics boards	≈ 10000

(*) The TriDAS system is designed to read 1 Mbyte data events up to 100 kHz level-1 trigger rate. In the first stage of implementation (Cost Book 9), the DAQ is scaled down (reduced number of RUs and FUs) to handle up to 75 kHz level 1 trigger rate

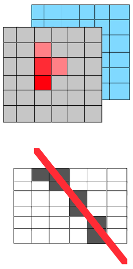
CMS trigger levels

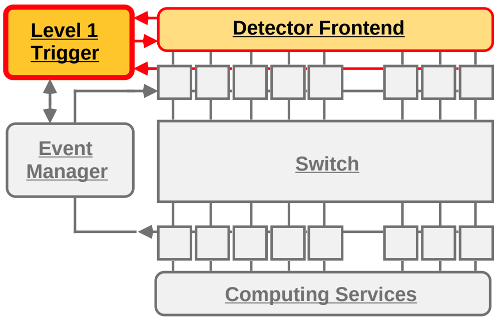
40 MHz



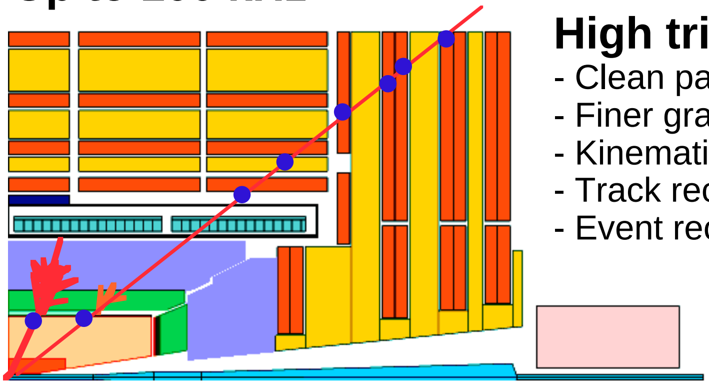
Level-1: Specialized processors

- Particle identification: high p_T electron, muon, jets, missing E_T
- Local pattern recognition and energy evaluation on prompt macro-granular information from calorimeter and muon detectors



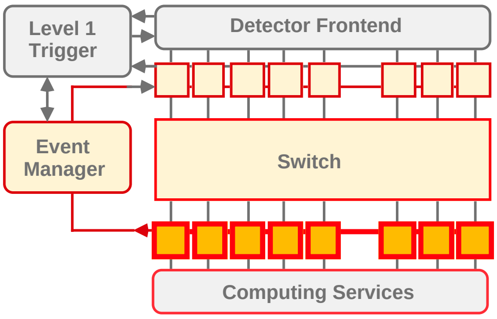


Up to 100 kHz

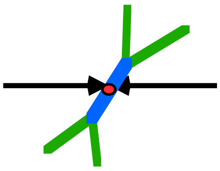


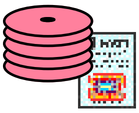
High trigger levels: CPU farms

- Clean particle signature
- Finer granularity precise measurement
- Kinematics. effective mass cuts and event topology
- Track reconstruction and detector matching
- Event reconstruction and analysis



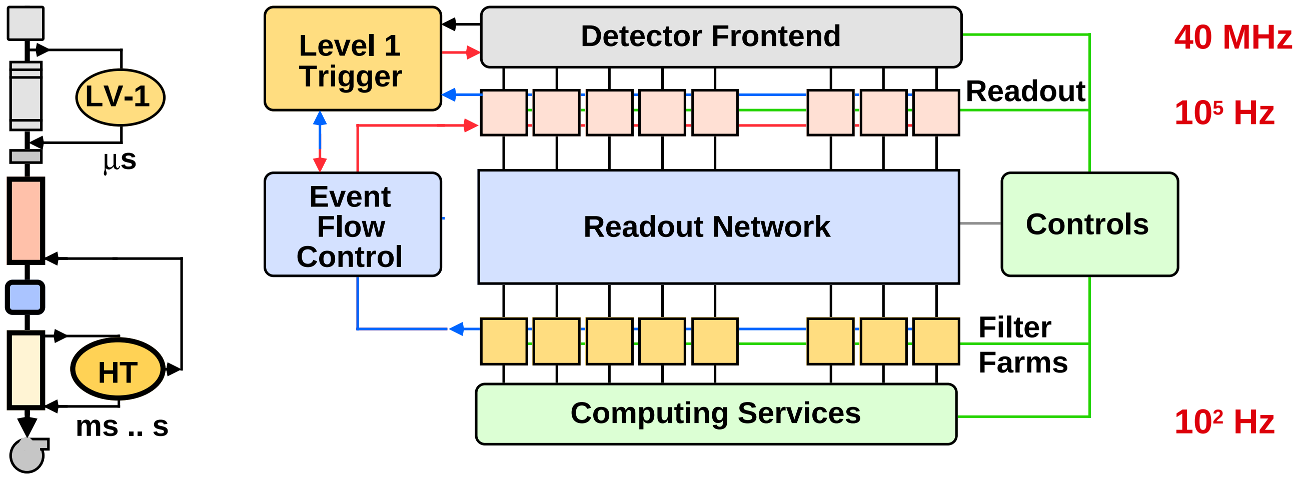
≈ 100 Hz





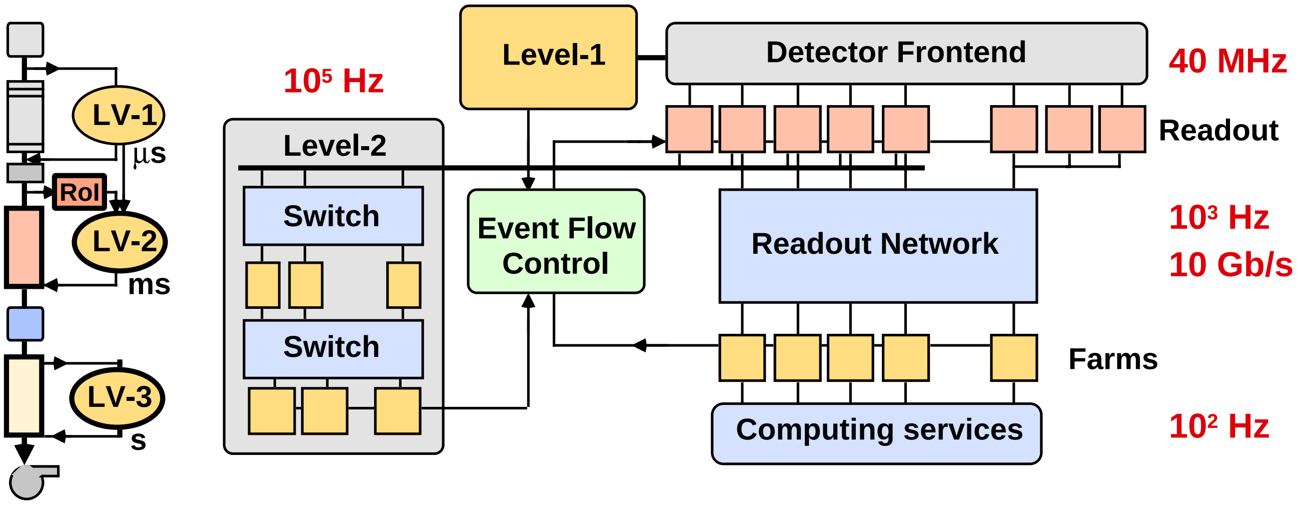
Structure options: physical levels

Two physical levels

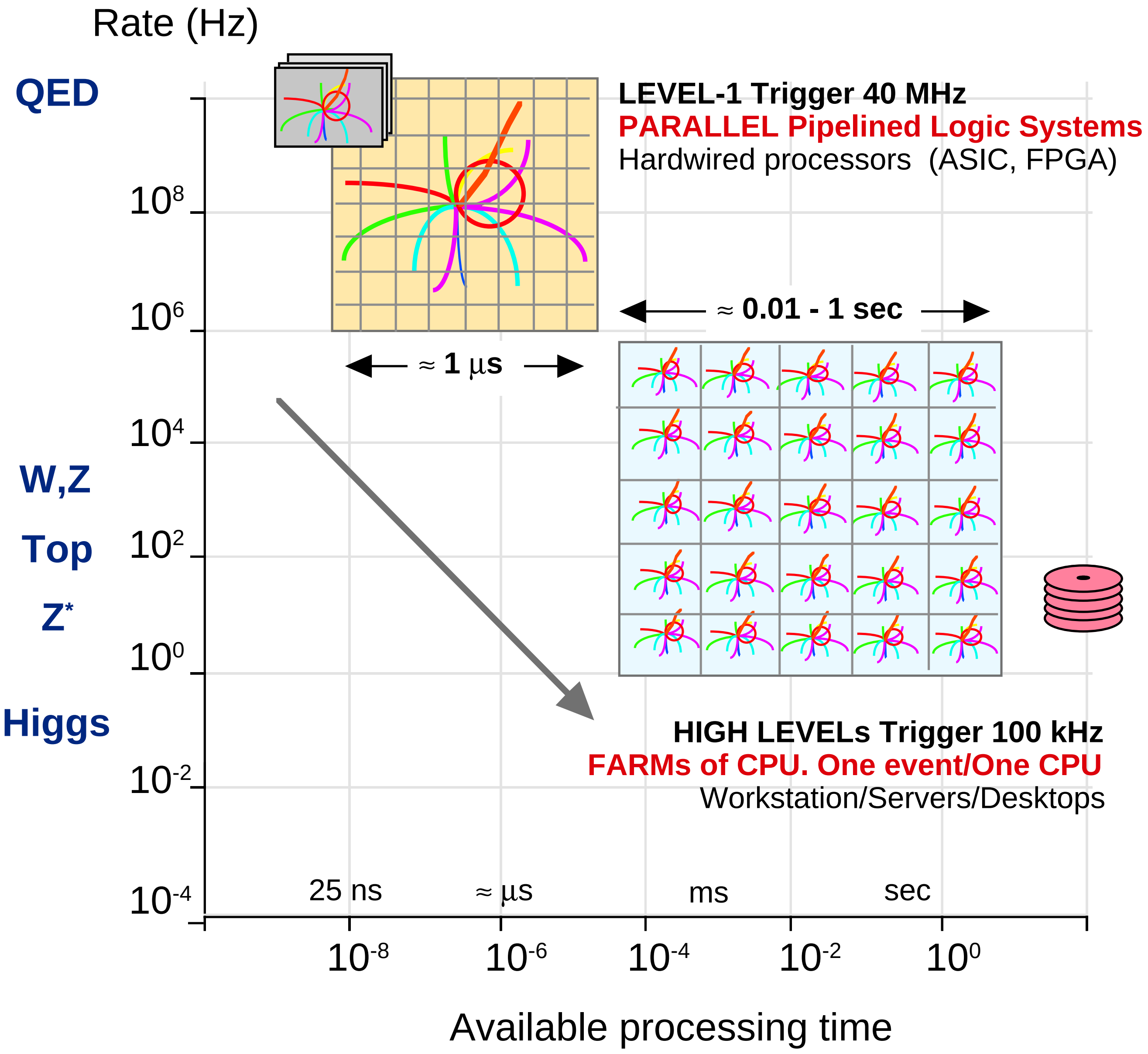


- Reduces the number of building blocks
- Exploits commodities: CPU, memory, links, networks
- Upgrades and scales with the machine performance.

Three physical levels



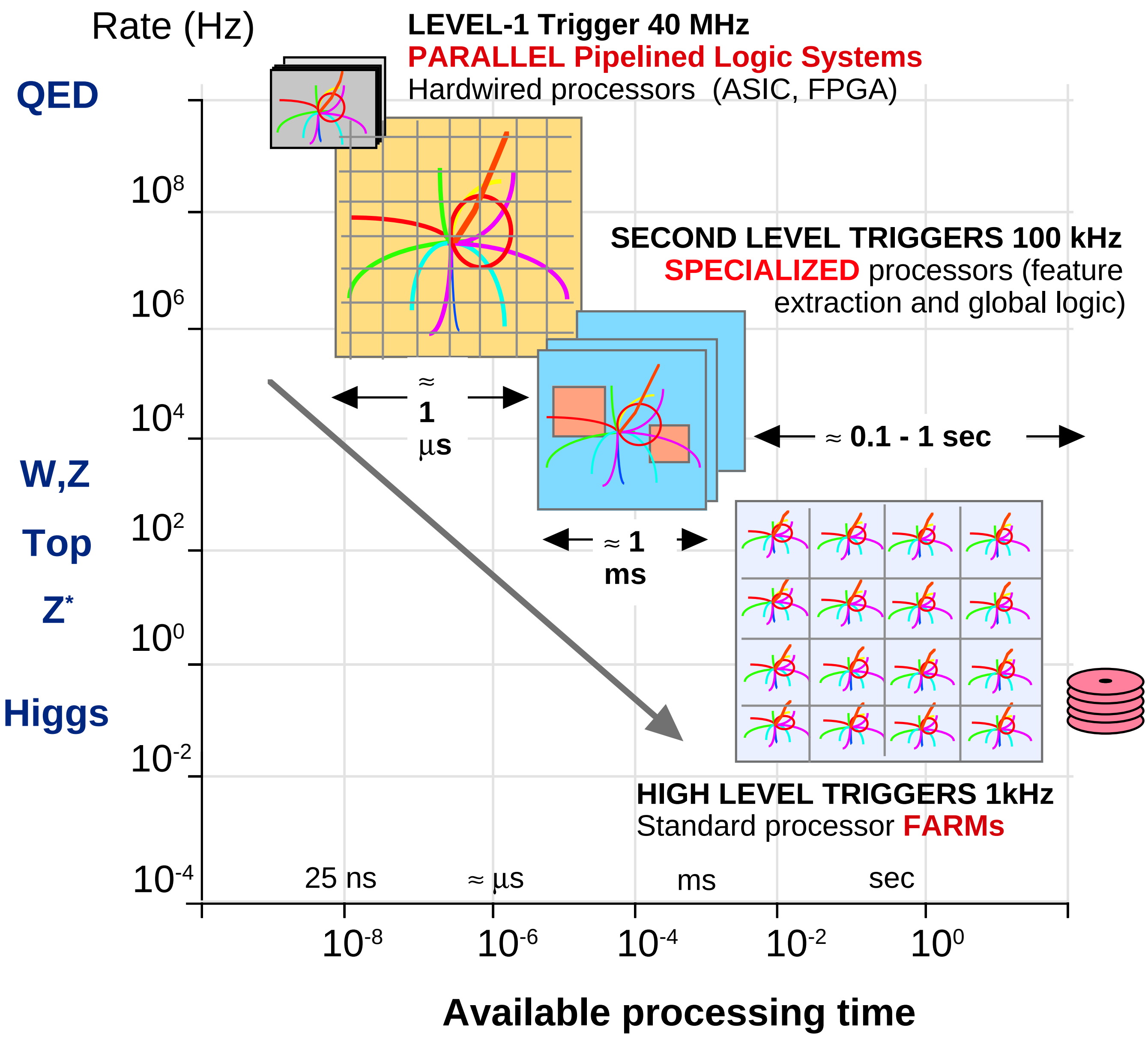
Event selection and computing stages (CMS)



LEVEL-1 Trigger 40 MHz
PARALLEL Pipelined Logic Systems
Hardwired processors (ASIC, FPGA)

HIGH LEVELs Trigger 100 kHz
FARMs of CPU. One event/One CPU
Workstation/Servers/Desktops

Event selection and computing stages (3)



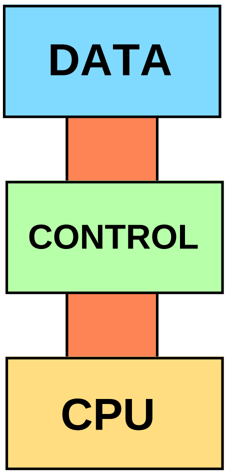
LEVEL-1 Trigger 40 MHz
PARALLEL Pipelined Logic Systems
Hardwired processors (ASIC, FPGA)

SECOND LEVEL TRIGGERS 100 kHz
SPECIALIZED processors (feature extraction and global logic)

HIGH LEVEL TRIGGERS 1kHz
Standard processor FARMs

High level triggers and readout structure

The DAQ structure is defined by the means used to make the data available to the trigger processor systems

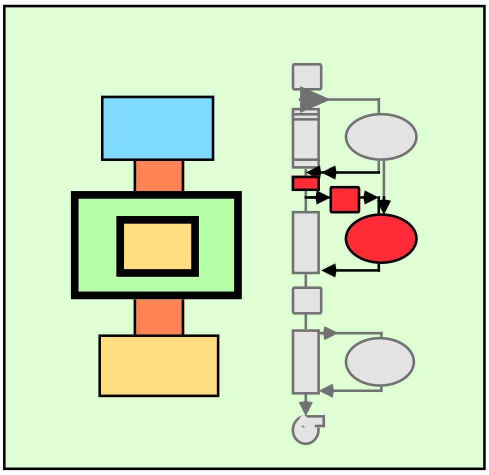


Readout Data

Bandwidth

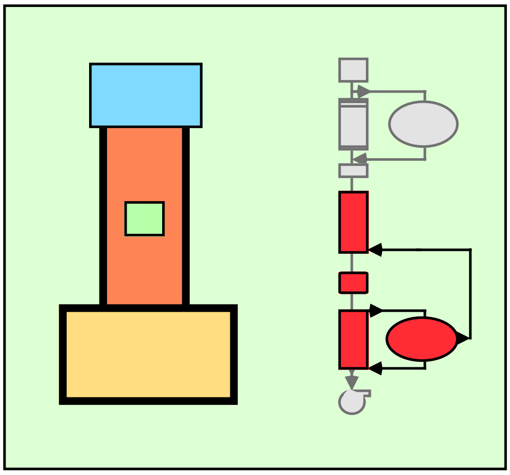
Data Access

Processing Unit(s)



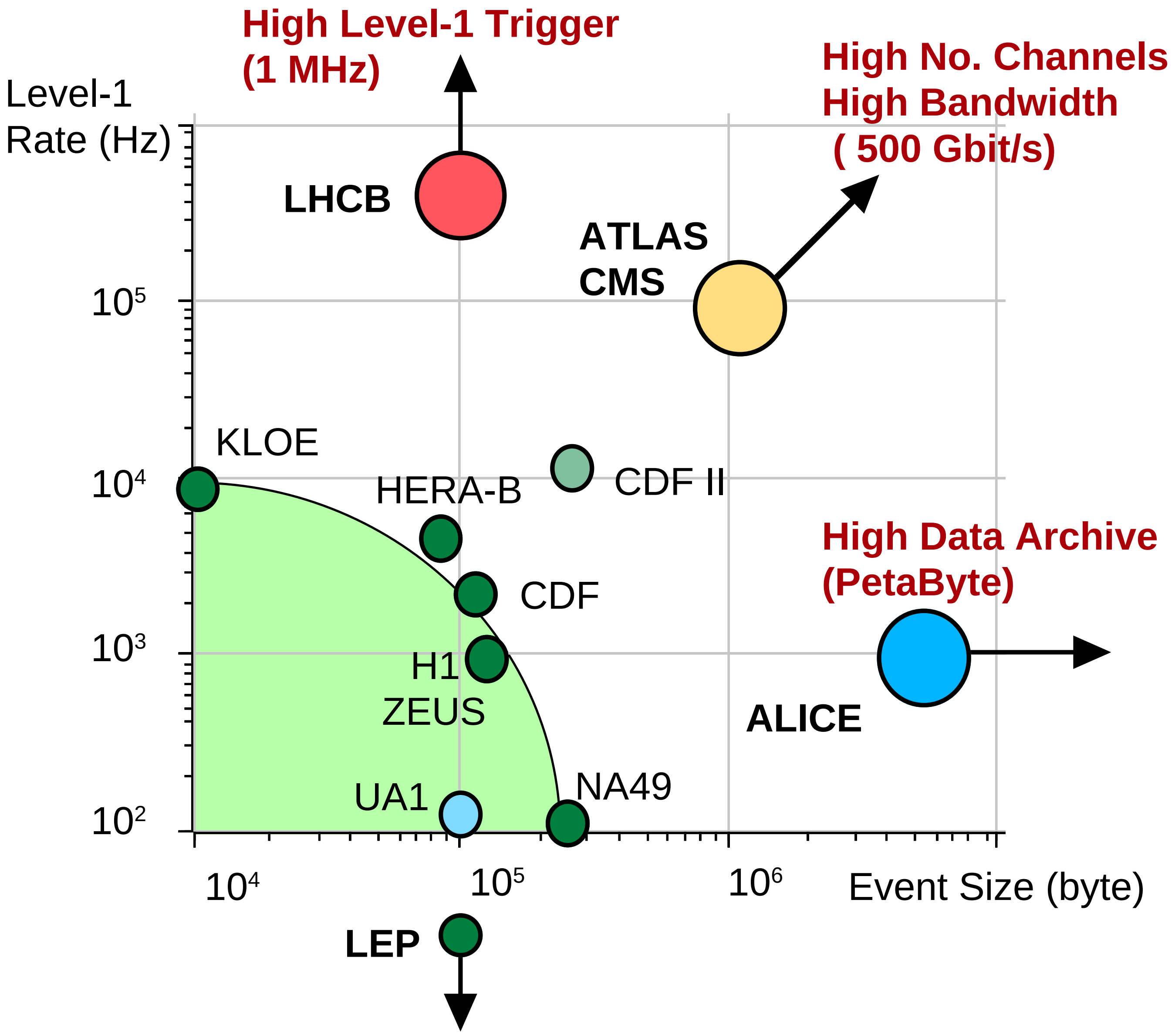
Invest in CONTROL logic and SPECIALIZED processors

CMS



Invest in BANDWIDTH and COMMERCIAL processors

Trigger and data acquisition trends



DAQ design status

DAQ main subsystems

Frontend logical model

Timing, Trigger and Control distribution

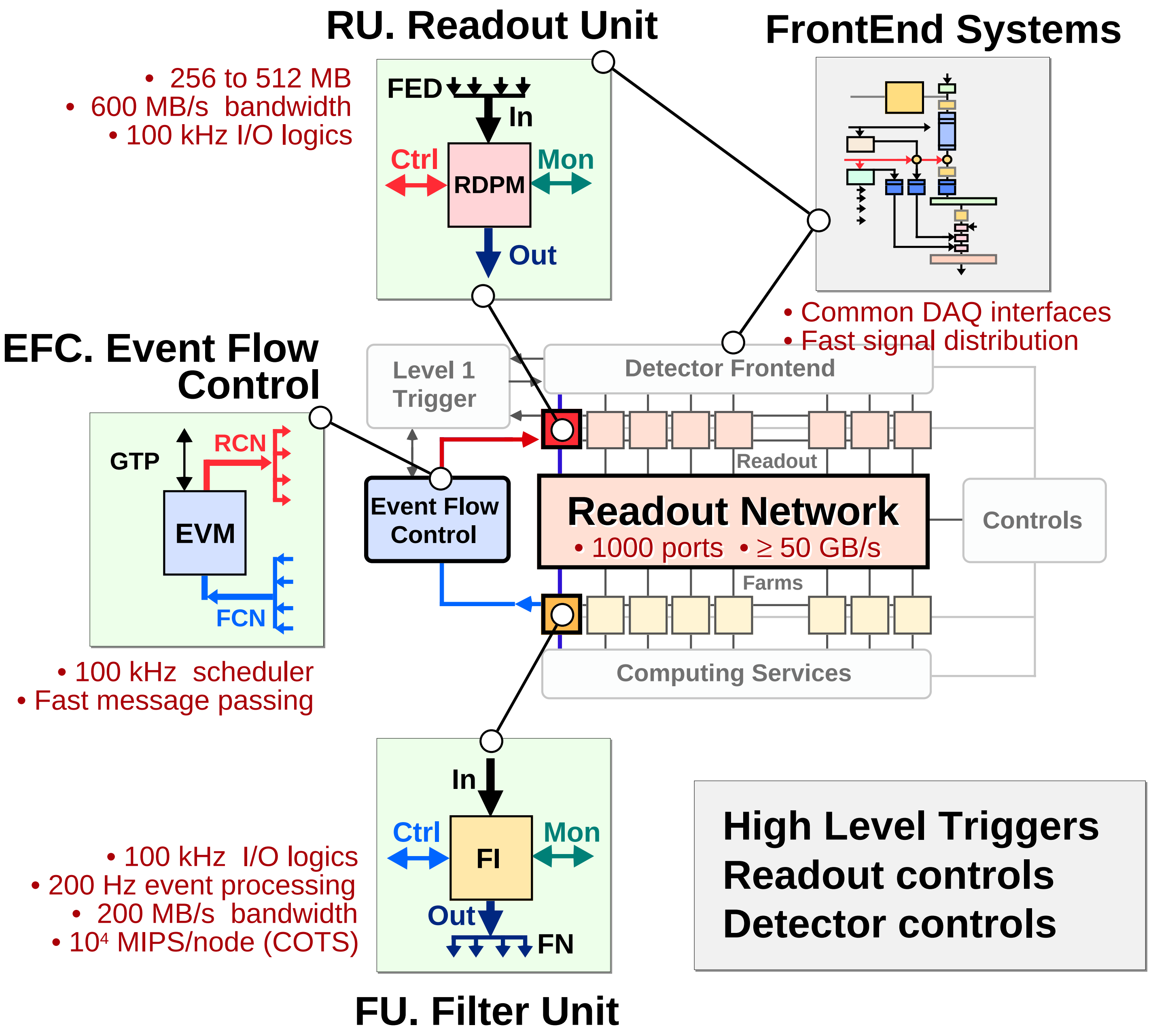
Readout Unit (RU): requirements

Filter Unit (FU): requirements

Event Flow Control (EFC): requirements

Event builder and high level triggers

DAQ main subsystems



RU. Readout Unit

- 256 to 512 MB
- 600 MB/s bandwidth
- 100 kHz I/O logics

FrontEnd Systems

- Common DAQ interfaces
- Fast signal distribution

EFC. Event Flow Control

- 100 kHz scheduler
- Fast message passing

- 100 kHz I/O logics
- 200 Hz event processing
- 200 MB/s bandwidth
- 10^4 MIPS/node (COTS)

FU. Filter Unit

High Level Triggers
Readout controls
Detector controls

RU	**Readout Unit**
FED	Front End Driver
RDPM	Readout Dual Port Memory
In	Data input port
Out	Data output port
Ctrl	Fast control port
Mon	Monitor port

EFC	**Event Flow Control**
EVM	Event Manager
GTP	Global Trigger Processor
RCN	Readout unit Control Network
FCN	Filter unit Control Network
FU	**Filter Unit**
FI	Filter Interface
FN	Filter Node

DAQ requirement summary

DAQ subsystems:

FrontEnd-DAQ interface

- Interface logical model
- Fast controls (TTC and TTS)

Readout Unit. RU

- FrontEnd Driver (FED) readout
- Readout Dual Port Memory (RDPM)
- Multi event buffering (256 MB) x 512 units
- Event builder data source (200-400 MB/s) bandwidth

Filter Unit. FU

- Filter Interface (FI). Event builder protocols
- Level-2 input event 200 Hz rate (≈ 100 MB/s)
- Level-3 input event 30 Hz rate (≈ 30 MB/s)
- CPU farm (≈ 10000 MIPS) x 512 units
- Farm output ≈ 1 Hz

Event Flow Control. EFC

- Assign events to destinations
- Assign Level-1 Triggers to DPM Event-IDs
- Receive Level-2/3 requests, forward to RU's
- Throttle Level-1 Trigger
- Operate at ≤ 100 kHz Event Rate

Readout Network

- 512 x 512 switch based event builder
- Bisection effective bandwidth ≥ 50 GB/s

Related issues:

High Level Triggers

- Physics channels, role of detector data
- Algorithms and processing power

Readout controls

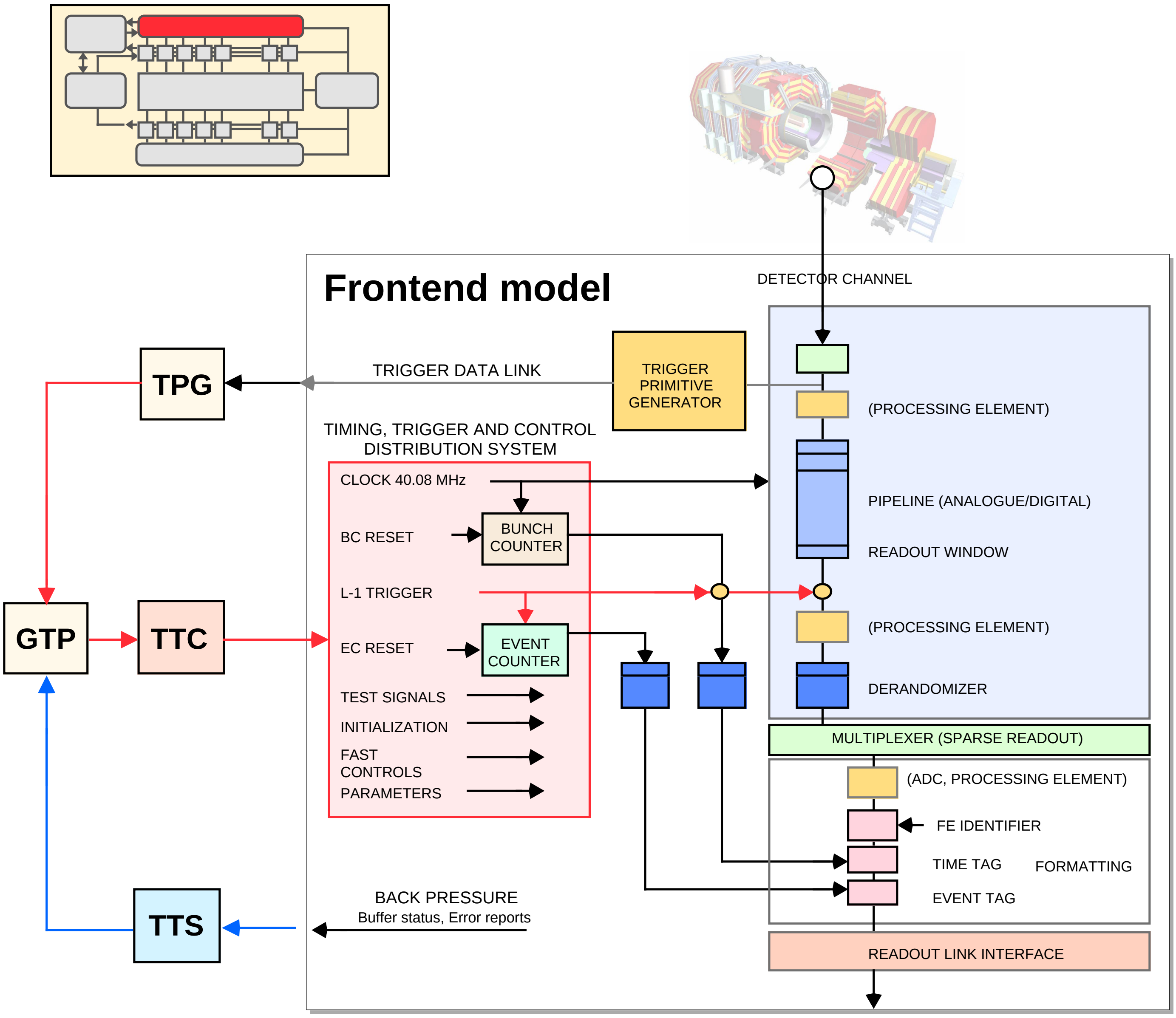
Detector controls

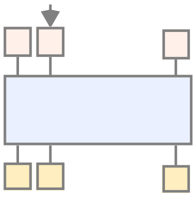
Parameter optimization

- Choice of technologies
- System behaviour extrapolation
- Protocols and functions evaluation

Computing and communication software

Frontend logical model

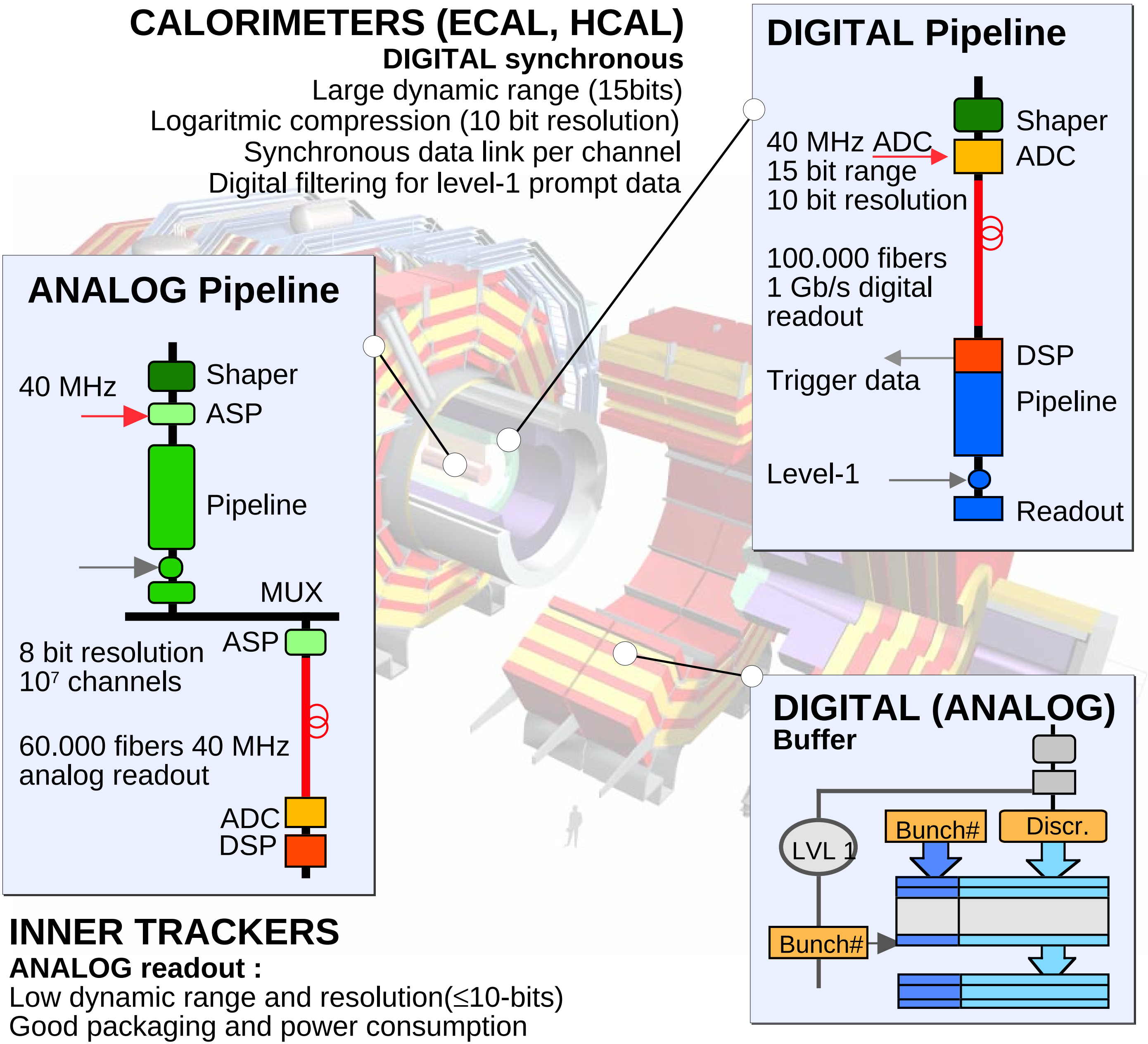




TPG	Trigger Primitive Generator
TTC	Timing, Trigger and Control
TTS	Trigger Throttle System
GTP	Global Trigger Processor

At the front-end, the detailed implementation of the various analogue and digital readout systems is governed by sub-detector-specific requirements which lead to the different configurations of VLSI and discrete electronics. However, from the point of view of the DAQ system much of the basic functionality of the front-ends is common to all the sub-detectors. All the readouts incorporate several stages of pipelining, derandomising, zero-suppression and signal processing, and are driven by the 40.08 MHz machine clock, Level-1 trigger accept and bunch counter synchronization signals distributed by the timing, trigger and control (TTC) system.

Frontend physical structures



CALORIMETERS (ECAL, HCAL)

DIGITAL synchronous

Large dynamic range (15bits)
Logaritmic compression (10 bit resolution)
Synchronous data link per channel
Digital filtering for level-1 prompt data

INNER TRACKERS

ANALOG readout :

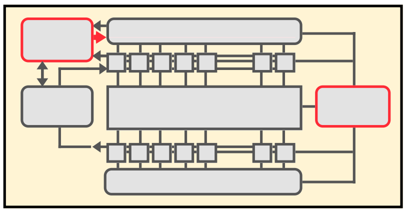
Low dynamic range and resolution(≤10-bits)
Good packaging and power consumption
High number of channel > 10^6
detector: inner tracker, pixel, preshower

MUON (RPC,DT,CSC)

DIGITAL asynchronous

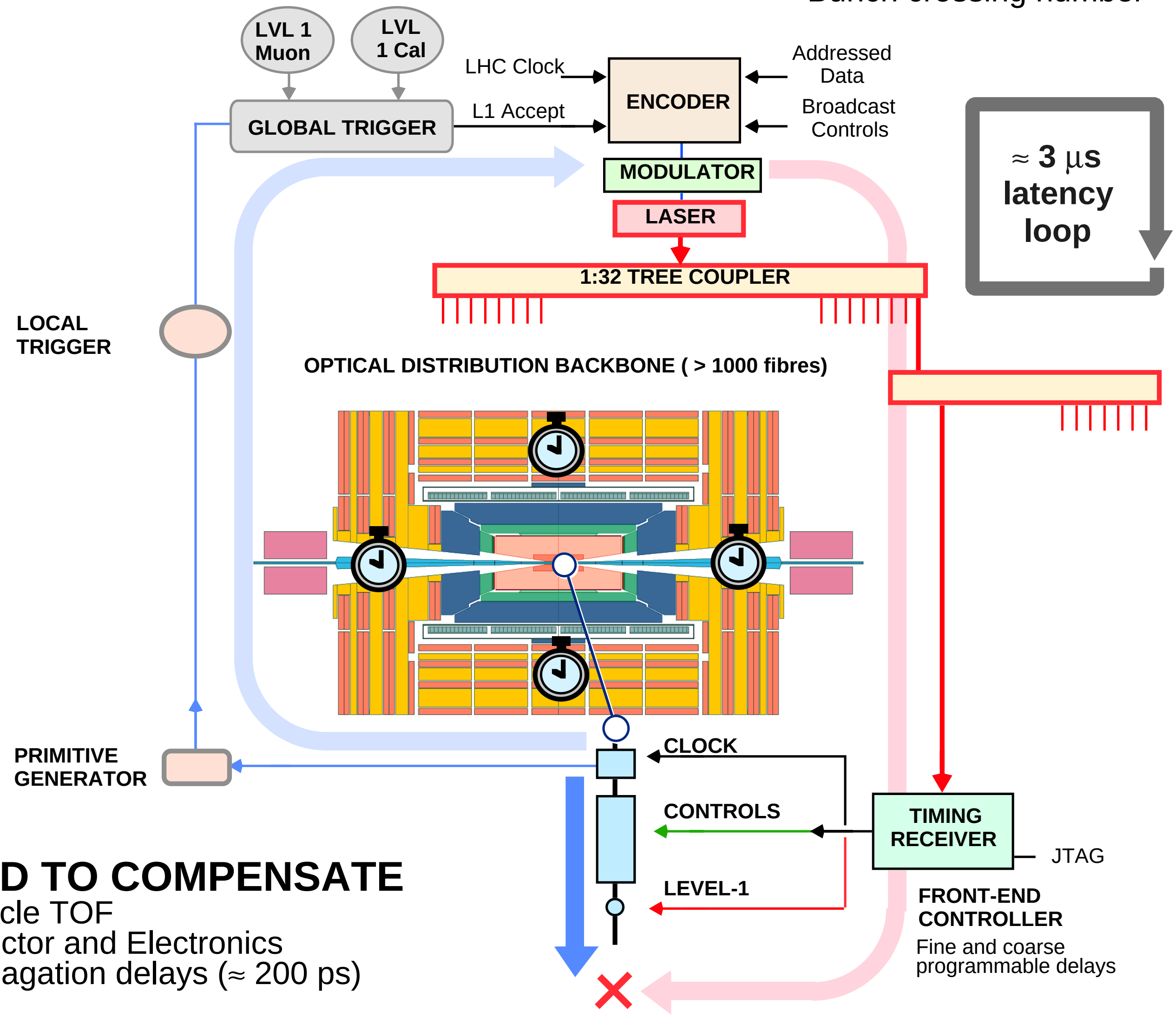
Low latency detector.
Time tag identification system.
detector:s: muon, trigger, (pixel analog memory)

Timing, Trigger and Control distribution



TTC system

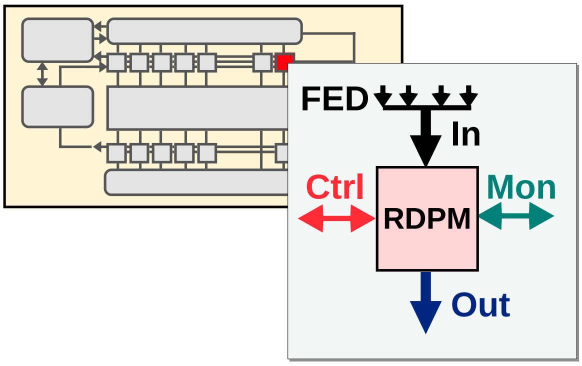
NEED TO DISTRIBUTE
- LHC clock
- Trigger-1 acceptance
- Control signals
- Addressable data
- Bunch crossing number



NEED TO COMPENSATE
- Particle TOF
- Detector and Electronics
- Propagation delays (≈ 200 ps)

The correct measurement of event data requires a precise timing broadcast system to synchronize the front-end electronics distributed around the apparatus with sub-nanosecond accuracy.
In CMS an optical network will be used to broadcast the level-1 trigger acceptance and addressable control information together with the 40 MHz clock. The receivers will be located in the front-end electronics systems and in the readout boards
The relative delay of each electronics subsystem will be programmable in units of the bunch crossing interval at the level of the pipeline control and in sub-nanosecond units at the level of the local timing and control station. Global timing calibration will be accomplished by exploiting the correlation between event signals and the bunch structure of the LHC machine.
The TC system will also be used at other levels in the readout network, such as the control of the final stage of the event builder, where synchronous message broadcasting is required.

Readout Unit (RU): requirements



IN (link/bus)	**400 MB/s**
OUT (link)	**200 MB/s**
Ctrl (network)	**≥10 MB/s**
Mon (network)	**≥10 MB/s**
MEM	**≥256 MB**

The Readout Unit (RU) is the logical system connecting a detector readout partition to one or more ports of the event builder network. In the current design the maximum number of RUs is 512. In the simple assumption of full event building 'at once', each RU reads and sends the same amount of data per event (2KByte in average). In the case of incremental event building protocols, driven by High Level Triggers (HLT), it may be necessary to configure the RU and FED partitions in such a way that all the switch ports contribute with the same average data rate even if the event sending rate is different (because the event builder process can be aborted by HLT decision before competition).
The RU main functions are:
- Readout of FrontEnd Driver modules on reception of a level-1 trigger signal generated by EVM and distribute by the Readout Control Network (RCN). The FEDs may be an integral part of the RU physical system (e.g. embedded DDUs) or a remote system linked to RDPM via a dedicated interconnection.
- Buffer event data into internal memory at 'address' communicated by EVM via RCN. The free data memory and pointer resources are handled independently in each RU while the event data handles (event ID) are unique and controlled centrally be the EVM.
- Send event data fragments to event builder via output data link according to the EVM commands.
- Perform event data flow monitoring and error reporting tasks during data taking. Operate as stand alone data acquisition system for test and maintenance.

The RU communication and processing systems consist of 4 ports (Input, Output, Control and Monitor) and of 2 subsystems (FrontEnd Driver (FED) and Readout Dual Port Memory (RDPM))

RU Ports :

INPUT - FED event fragment readout (n_{FED}=1 to 8)
- Input bandwidth up to 400 MByte/s
- Maximum event rate of 100 kHz. Event fragment sizes varying from 400 to 4000 bytes

CONTROL - Input event commands (≈100 kHz 32 bit messages)
- Output event commands (≈ 300 kHz 64 bit messages)
- Status output (few kHz 32 bit messages)

MONITOR - Full access to the module internal registers and memories.
- Alternative Data Input and Output ports and Data snap Shot during readout

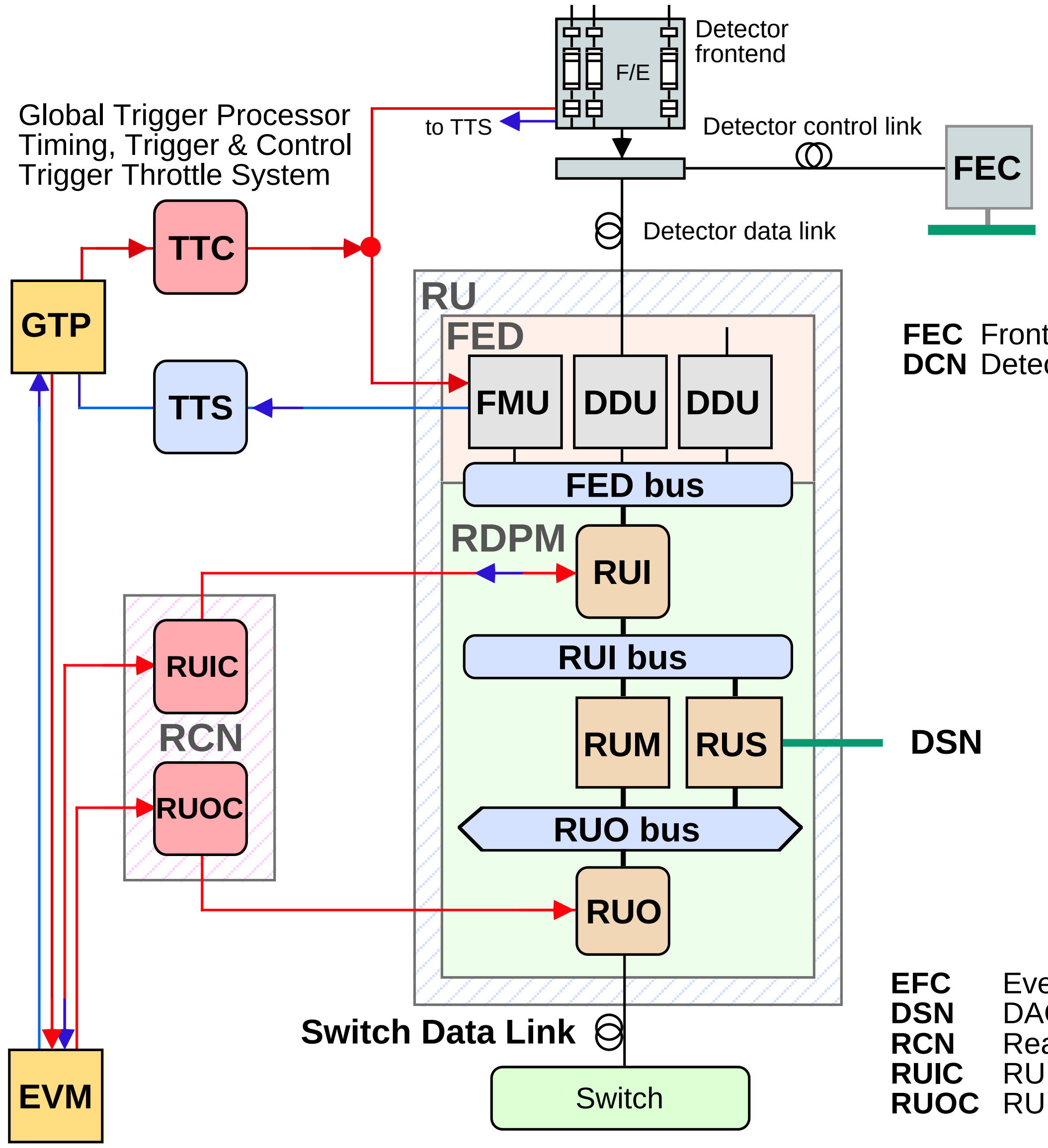
OUTPUT - 1 to n Output data links (n=1..4)
- Output bandwidth up to 200 Mbytes/s. (Lower ≈ 100 MB/s with level-2 HTL)

RU subsystems:

FED The FrontEnd Driver is a system collecting and formatting detector data from multiple Detector Depending Units (DDU) on response of a level 1 trigger.

RDPM The Readout Dual Port Memory is an autonomous data acquisition system reading one or more FEDs, buffering events and on request driving the output into the EVB network

RU: system architecture



GTP	Global Trigger Processor
TTC	Timing, Trigger & Control
TTS	Trigger Throttle System

FEC	FrontEnd Control
DCN	Detector Control Network

EFC	Event Flow Control
DSN	DAQ Services Network
RCN	Readout Control Network
RUIC	RU Input Controller
RUOC	RU Output Controller

The RU internal functionality is broken down as follows:

FMU Fast Monitoring Unit. FMU processes fast status and control signals (Level-1, error, busy..)

DDU Detector Dependent Unit. The DDU reads, formats and buffers the detector data

FED bus It is the interconnection between Detector Dependent Units (DDU) and RU Input (RUI)

RUI Readout Unit Input. On reception of Level-1 via RUIC interface, the RU memory resources are allocated and DDUs (FED) readout is performed under RUI control. RUI may perform higher level data handling function as well (check, simulation, formatting)

RUI bus It is the interface between RUI and RUM input data memory bus

RUM Readout Unit Memory. Data memory with dual access and event pointer/tagging logics

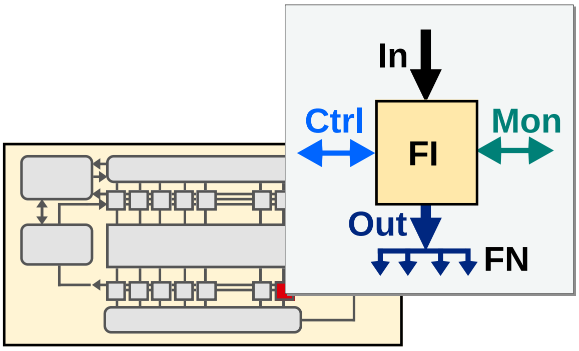
RUO Readout Unit Output. RUO executes EVM commands (RUOC interface) and initializes and controls data transfer to the switch according to the event builder protocols

RUO bus It is the interconnection between RU data memory and output data link

RUS RU Supervisor. Processor running high level software dedicated to RU initialization, auto-test, readout control and monitoring tasks (snap shot, synchronization, etc.)

Switch data link It is the interface to the switch input port (e.g. ATM, FCS, Ethernet ..)

Filter Unit (FU): requirements



IN (link)	**200 MB/s**
OUT (bus/link)	**200 MB/s**
Ctrl (network)	**≥10 MB/s**
Mon (network)	**≥10 MB/s**
MEM	**≥256 MB**
CPU	**~ 10 TeraOPS**

The Filter Unit (FU) is the logical system connecting one or more switch output ports to a processor system (Farm Node). At present 512 FUs are foreseen. The main FU functions are:
- Communicate with the EVM the status of internal resources (event buffers and CPUs). Send requests for sub-events to the EVM (a sub-event is the collection of data from a set of RUs)
- Read multiple event fragments coming from the switch port(s) and build events data buffers as required by the High Level Trigger (HLT) algorithm running in the Farm Node CPUs. The process is logically simultaneous for multiple sub-events.
- Get requests for more data or for deletion of event data from running CPUs.

The FU communication and processing systems consist of 4 ports (Input, Output, Control and Monitor) and of 2 subsystems (Filter Interface and Farm Node))

FU Ports:

INPUT - 1 to n data links (n=1..4). Bandwidth of 100 MB/s with incremental level trigger
E.g. level-2 event size ≈ 300 kB x 200 Hz. Higher levels event size ≈ 700 kB x 20 Hz
(200 MB/s if reading full 1 MB events)

CONTROL - OUTPUT Filter commands (few kHz 64 bit messages)
- Reset and set mode commands

MONITOR - Full access to the module internal data and control registers and memories.
- Alternative Data Input and Output ports and Data snap Shot during readout

OUTPUT - Interface with the CPU farm internal memory
- Bandwidth up to 100 MB/s

FU subsystems:

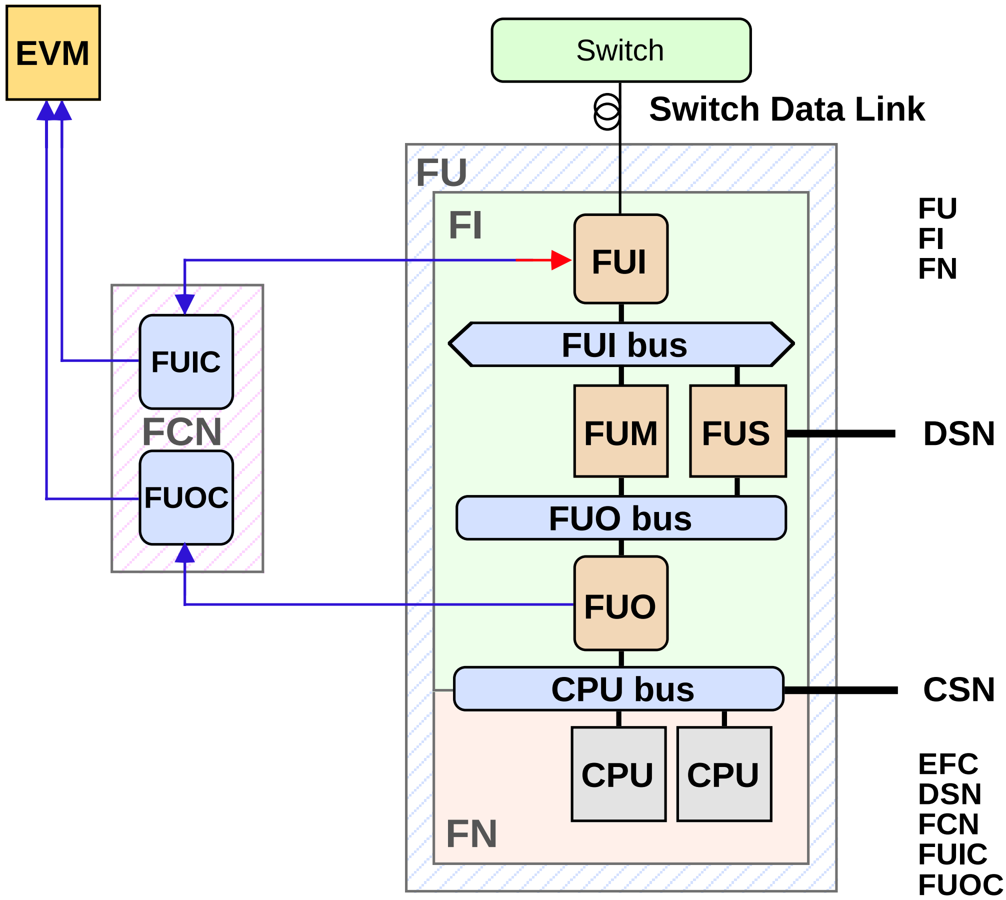
Filter INTERFACE (FI). It interfaces the DAQ event builder to a processor farm.
It reads multiple event fragments coming from the switch. It builds complete (or almost) events buffers and data are made available to processor via its I/O channel. The FI provides the communication path between the farm filter node processes and the event builder components (EVM, RUs, GTP..)

N is the number of switch output ports		(eg.N=1000)
Event fragment byte size	≈ 10^6 /N Byte	(1000 Byte)
Event fragment maximum rate	≈ 100 kHz	(100 MByte/s)
Full event processor input rate	≈ 10^6/N Hz	(100 Hz)

FARM NODE (FN). It is a commercial processor servers or desktops with a fast Input channel to read event data and a generic output interface to a data server system

Farm node processing power	≈ 10^7/N MIPS	(10000 MIPS)

FU: system architecture



FU	**Filter Unit**
FI	Filter Interface
FN	Filter Node

EFC	**Event Flow Control**
DSN	DAQ Services Network
FCN	Farm Control Network
FUIC	FU Input Controller
FUOC	FU Output Controller
CSN	Computing&Services Network

The FU internal functionality is broken down as follows:

FUI Filter Unit Input. Interface to the switch output port. The FUI includes the logic to handle the internal buffer resources, to communicate their status to EVM and to read and assemble the event fragments coming from the switch port interface.

FUI bus It is the interface between the switch port interface and FU data memory

FUM Filter Unit Memory. Data memory, dual access and event pointer/tagging logics

FUO Filter Unit Output. FUO communicates to EVM the status of event processing and data requests

FUO bus It is the interconnection between FU data memory and Farm CPU bus

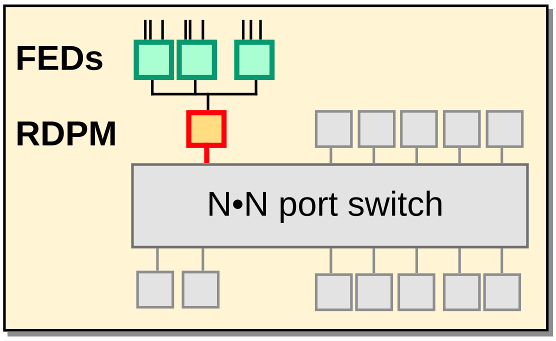
FUS FU Supervisor. Processor running high level software dedicated to FU initialization, auto-test, readout control and monitoring tasks (snap shot, synchronization, etc.), It can be one of farm processors

CPU bus Data I-O bus/channel specific to the Processor system

FCN Filter Unit I/O Networks. They are the communication and concentrator systems of all DAQ (fast) messages generated by the SFI (farm) units, FCN is interfaced to the event manager communicating the status of each processing resource running in the farms (the Farm Status collection system may be implemented by the switch itself or by a separate message collection system)

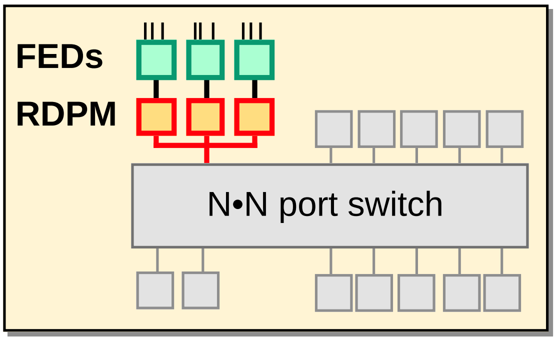
CSN The Computing Services include local/global data staging/archiving systems and network services. The maximum global throughput for mass storage is of the order of few 100 MByte/s. They provide also the (slow) control and monitor of the farm processors (downloading, diagnosing, monitoring etc.)

RU-FU architectural features



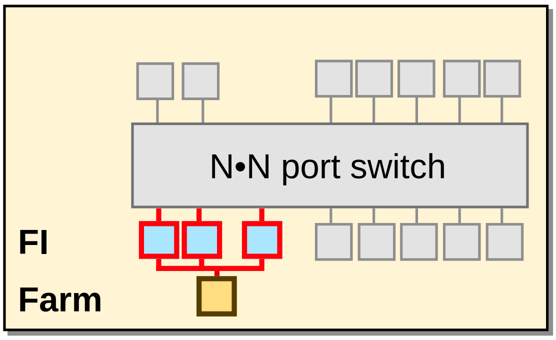
Scaling No. RDPMs

RU configuration with multiple FED systems. To be used to readout low occupancy detector partitions or at low luminosity with a reduced size switch



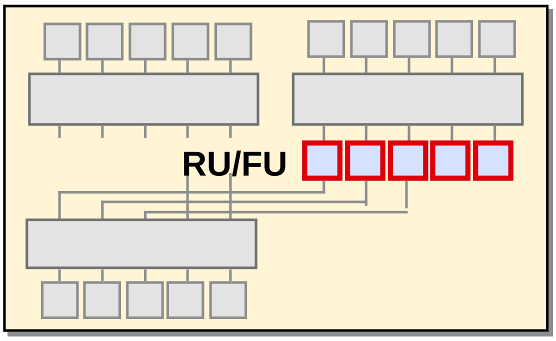
Scaling RU Input bandwidth

RU configuration with input bandwidth higher than output. To be used to readout detector partitions contributing to high level trigger (e.g. tracker) where a large fraction of events is expected to be rejected and a switch input data balance is required.
This configuration can be used to scale DAQ at switch port input level (reduced switch size etc.)



Scaling FU input ports

FU configuration with multiple switch port interfaces associate to a single farm system. Multiple interfaces can be linked to the same FU memory or each link interface may have its own FUM resource.
This scalability can be used in a descoping scenario where readout capacity is preserved and only the farm nodes are reduced

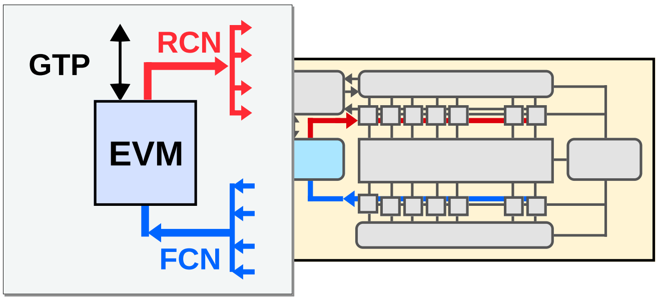


Switch fabrics : middle layer buffers

RU-FU event handling logics are very similar.
RUs can be used also as intermediate event builder buffers (super fragments). The characterization of RDPM/FI is implemented by the RUI and FUI systems

Event Flow Control (EFC): requirements

EFC	**Event Flow Control**
EVM	Event Manager
RCN	Readout Control Network
FCN	Farm Control Network

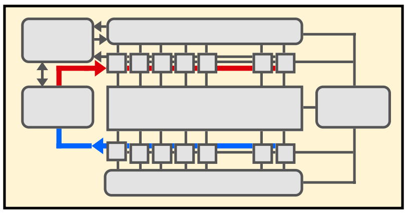


The Event Flow Control (EFC) is responsible of the control, execution and synchronization of the event building protocols. In the current baseline design the EFC consists of a processing unit (the Event Manager EVM) and two networks, one collecting the status of all Filter Units (Filter Control Network FCN) and the second distributing the readout and event building commands to the Readout Units (Readout Control Network).

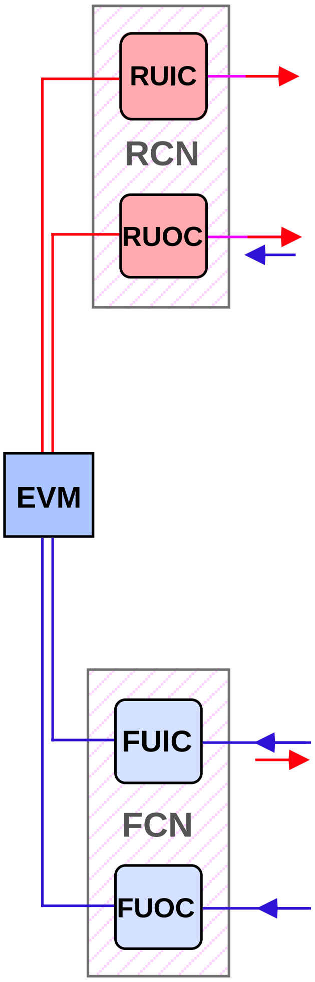
The EVM performs the following functions:

- Receive and book Level-1 Triggers from Global Trigger Processor
- Administrate the generation of the Event Identifiers (EID). The EID is the data handle used by RU to address an event data. The EID is unique for each event stored in the RUs
- Distribute Level-1 Trigger commands to RU with assigned EIDs
  Examples of commands: OPEN event, READ, CLOSE and their combinations
  Command parameters: Event Identifier, Event type
- Receive from FUs the High Level Trigger (HLT) requests and distribute the event building commands to RUs. E.g. Event Filter resource ID, RU set, EID
- Monitor readout buffer levels and throttle Level-1 Trigger
- Operate at ≤ 100 kHz Event Rate

EFC: control networks



The EVM uses two (logical) networks to control and monitor the data transport from the RUs into the FUs via the switch: the Readout Control and the Filter Control Networks (RCN, FCN)



RCN. Readout Control Network

RUIC. Readout Unit Input Controller
Broadcast read event message (1 to N≈512):
- Level-1 trigger and Event ID (100 kHz 32 bit messages)
- Reset signal(s) (up to 100 kHz)

RUOC. Readout Unit Output Controller
Distribute 1 to x(selectable set) ≈300 kHz 32 bit EVM commands:
- Associate Event (#ID) to Filter Unit resource (#CPU)
- Send event data to output port
- Flush event pointer
- Set mode/Test/Reset etc.

Collect (merge N≈512 to 1) RU resource status
- Buffer occupancy (100 kHz strobe)
- Fault messages (few kHz 64 bit)

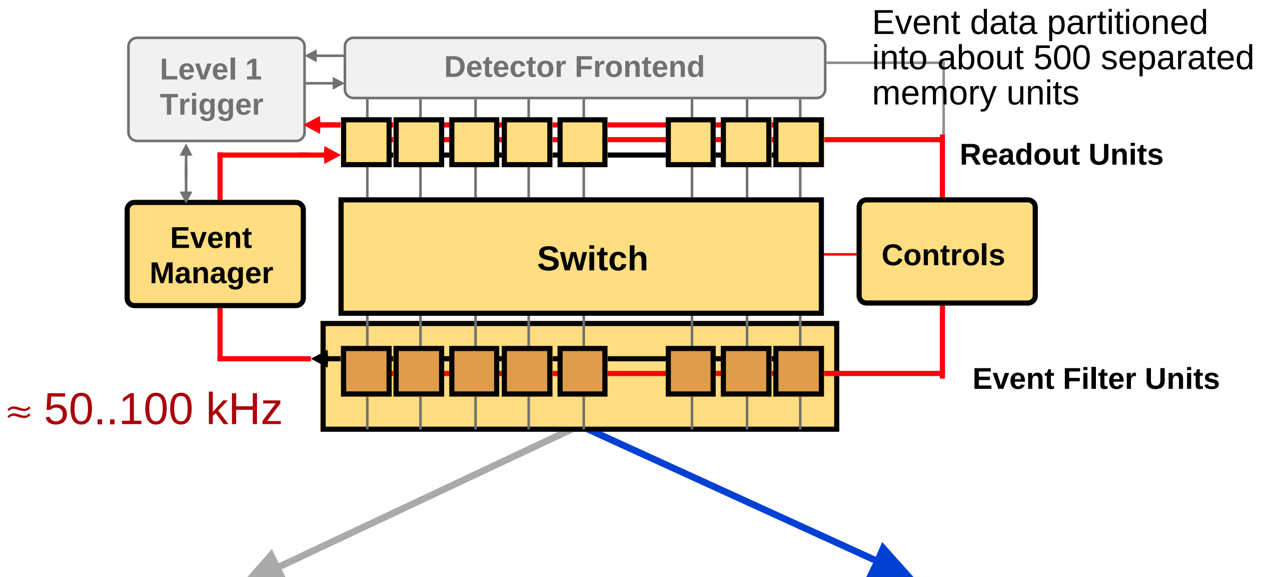
FCN. Filter Control Network

FUIC. Filter Unit Input Controller
Collect Filter Unit resource status (N≈512 to 1)
- FU(FI) free event buffers (≈300 kHz 32 bit commands)
- Event assembly fault messages (few kHz 64 bit)

Distribute FU control messages
- Cross check information (e.g. #ID next event)
- Set mode/Test/Reset

FUOC. Filter Unit Output Controller
Collect Filter Unit messages:
- High level messages (fault & trigger back pressure ..)
- Event Builder Reset commands

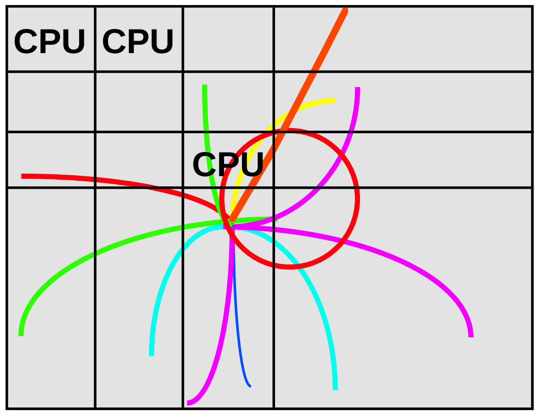
Event builder and high level triggers

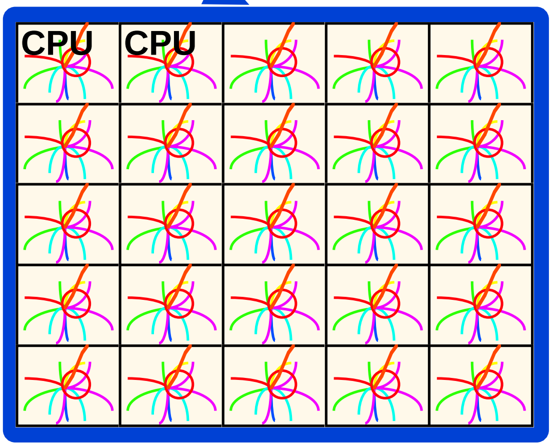


Event data partitioned into about 500 separated memory units

Readout Units

Event Filter Units





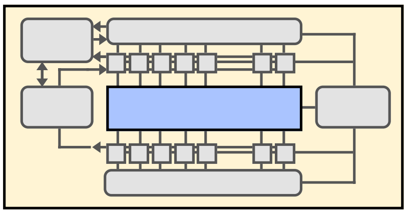
Massive parallel system
ONE event, ALL processors

- Low latency
- Complex I/O
- Parallel programming

Farm of processors
ONE event, ONE processor

- High latency (larger buffers)
- Simpler I/O
- Sequential programming

Event Builder : "switch" parameters



CMS DATA ACQUISITION parameters:

Average event size	Ev ≈ 1 Mbyte
Level-1 physics rate	L1 = 30 kHz
Switch efficiency	S_L ≈ 50% (safe factor)

FULL EVENT BUILDING at 30 kHz requires a NxN cross bar switch with a port speed = (Ev•L1) / (N•S_L) = 60•10^9/N B/s

Switch technologies (N•N) for 30 kHz, 1MB/Ev and 50% load

It is assumed that the technology allows the integration of a N•N switch with bisection bandwidth= N x port speed and non blocking cross-bar NxN

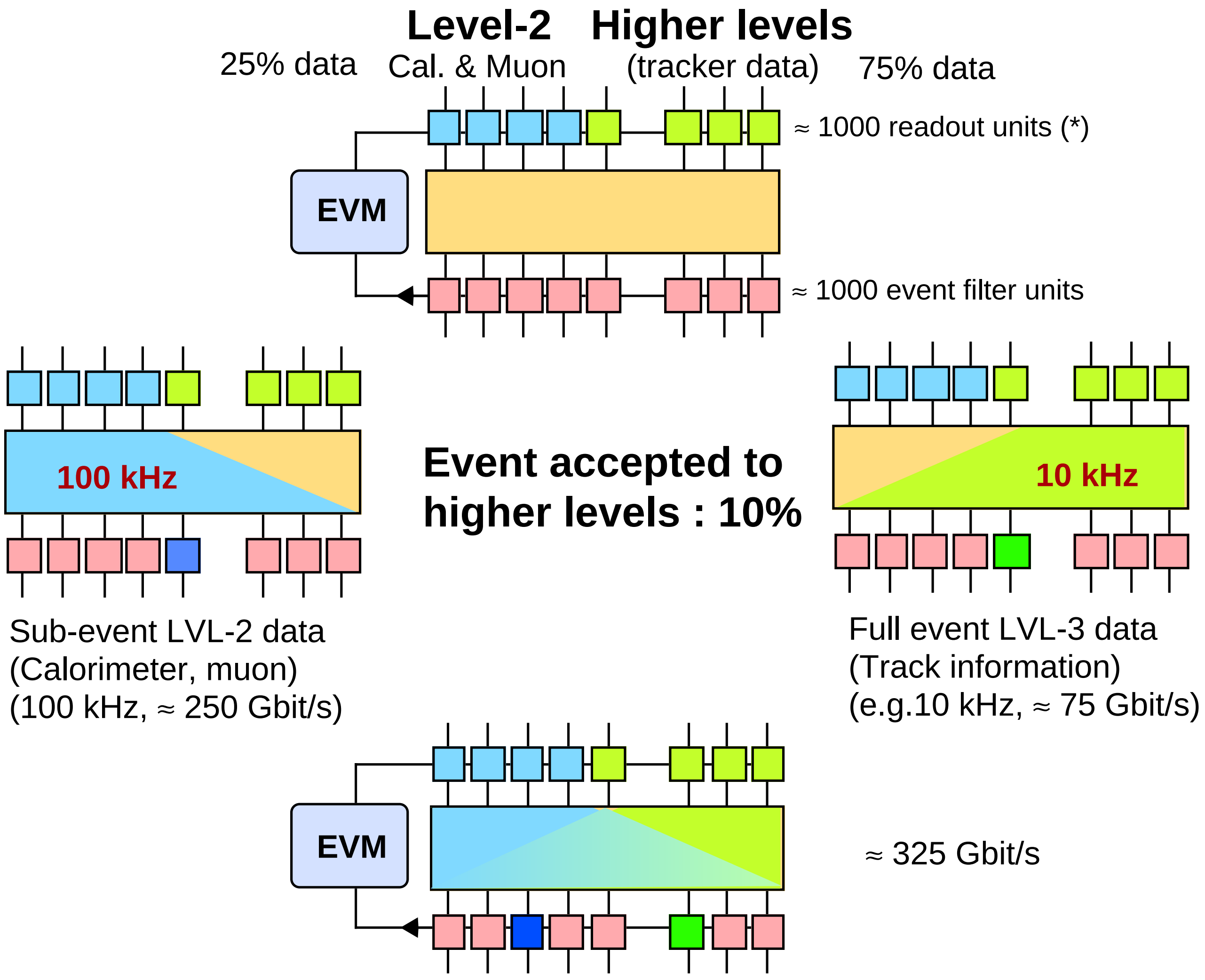
No. Ports	Required port speed		Technology	Link speed	
N = 1000	Ps ≈ 60	MB/s	ATM,GE, Myrinet	622	Mb/s
N = 500	Ps ≈ 100	MB/s	FCS,GE, Myrinet	1	Gb/s
N = 256	Ps ≈ 240	MB/s	ATM, Myrinet II	2.4	Gb/s
N = 64	Ps ≈ 1000	MB/s	GEx10	10	Gb/s

The CMS event builder network is designed to read 1 MByte events at 30 kHz Level-1.

It can be extended to read higher trigger rates (up to 100 kHz) implementing selective readout protocols driven by the HIGH LEVEL TRIGGERs running the filter farms

1) Event building by steps

Event is built in two (or more) steps under HTL control.
This allows to extend the the event building rate up to 100 kHz using a lower perfomance switch



Level-2 **Higher levels**

25% data Cal. & Muon (tracker data) 75% data

≈ 1000 readout units (*)

EVM

≈ 1000 event filter units

100 kHz

Event accepted to higher levels : 10%

10 kHz

Sub-event LVL-2 data
(Calorimeter, muon)
(100 kHz, ≈ 250 Gbit/s)

Full event LVL-3 data
(Track information)
(e.g.10 kHz, ≈ 75 Gbit/s)

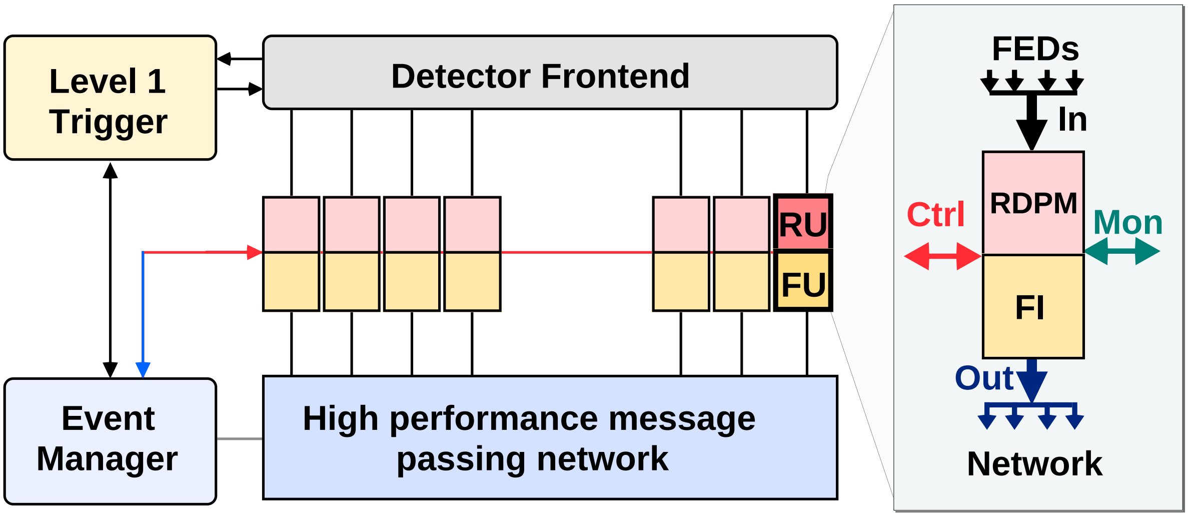
EVM

≈ 325 Gbit/s

Required switch bandwid =325 Gbit/s (e.g. Myrinet 512x512 at 60%)

The principle, is that for each event accepted by the Level-1 trigger, only a fraction of the detector output is transported into one or several processors, capable of making a refined analysis of the event based on this limited information. This is the "Level-2" decision on the event..
In order to achieve the data acquisition figure of 100 kHz event rate after the level-1 trigger (with limited switch bandwidth), the tracking data are not moved into the readout network until the associated event has passed the test of the high trigger levels based on the information from the other detectors. This operation (called virtual level-2) is expected to reduce the event rate (for the tracker data) by at least one order of magnitude.
(*). A more complex data balancing scheme is foreseen in order to have an equal sharing of the switch bandwidth between level-2 and level-3 sources.

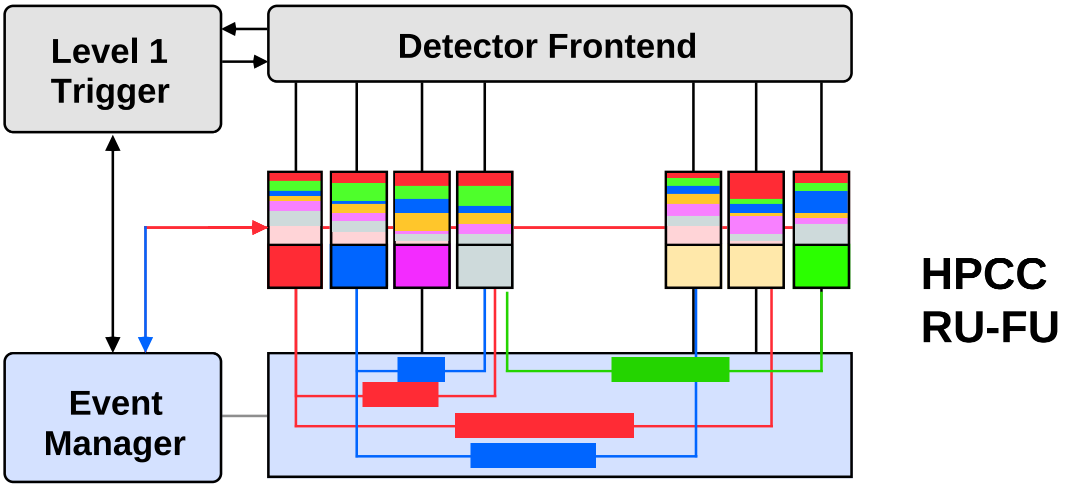
2) High Performance Computing and Communication event builder



Readout and Filter units are implemented by the same physical unit (RU-FU). The switch interconnects all RUFUs. Communication software allows transparent access between all RUFUs internal resources (readout buffers)

The event manager has the same event input tasks (e.g. LV1, EvID) and some broadcast functions (if these can not be implemented by the switch itself (e.g. open event, close event, reset etc.)

HPCC pull data mode EVB



Filter unit pulls data on HLT requests.

Each RUFU node controls one or more readout unit. Events are associated to RUFU processes either by an automatic (e.g. Ev#) or an external (e.g. EVM) event allocation system. According to the algorithm execution step the event data fragments (stored in the other RUFUs) are fetched through the switch by the RUFUs communication layer when needed.

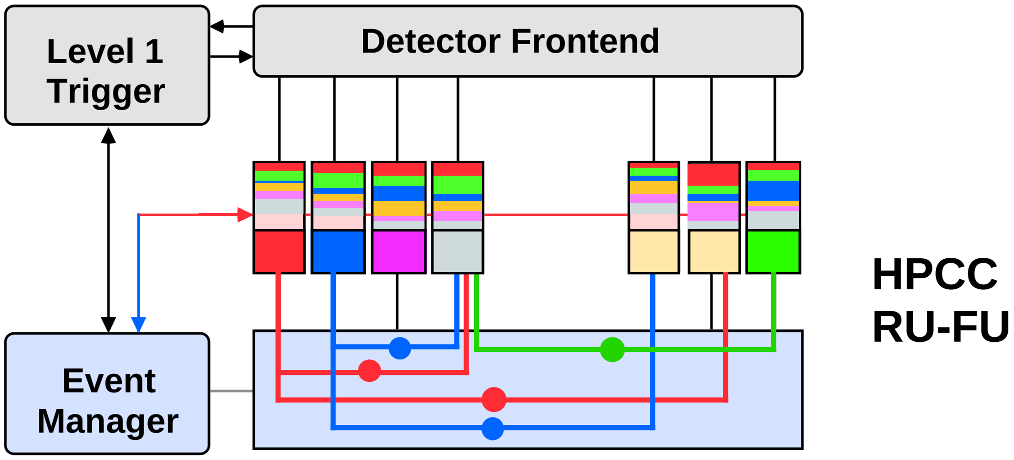
Required port bandwidh = .48 Gb/s (switch load 40%)

Pb = l1 (n1•M1 + f1•n2•M2 + f1•f2•n3•M3 etc..)

(e.g. n1=n2=200, M1=M2=10000, f1=.2, neglect other terms)

L1	DAQ Level-1 trigger rate	(100 kHz)
N	Total number of RUFU units	(512)
l1	FU trigger rate	(200 Hz)
Pb	Bandwidth (bits/s) required at each port	
n1	Number of messages to be exchanged for HLT level 2	
M1	Average size (bits) of message for HLT-2	
f1	Fraction of events passing HLT-2	
n2,M2,f2	The same for HLT-3 etc.	

HPCC OO paradigm EVB



Event filtering by OO paradigm

The same as before, but data are not transferred. Each 'event filter process' runs the analysis of one event. The data processing is executed by the processor resource of the RUFU holding the data. 'Methods' are invoked by fast message passing through the switch. This solution needs data fragments well mapped to detector subsystems and auto-contained information suitable for local reconstruction etc. etc.

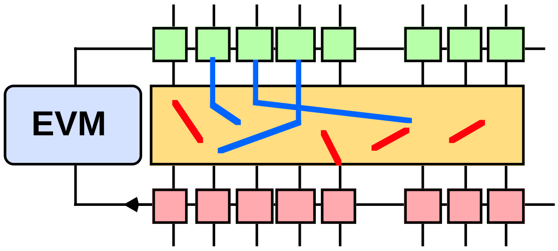
Required port bandwidh = .048 Gb/s (switch load 4%)

Mass storage. In both the cases the event selected are sent to mass storage either by the same switch (just a small amount of bandwidth is used) or by the standard network (e.g. GEthernet) which surely will interconnect all systems.
100 Hz (selected events) event building for mass storage requires an additional port bandwidth of few Mb/s (switch load of 0.2%)

Pb = l1 (n1•M1 + f1•n2•M2 + f1•f2•n3•M3 etc..)
E.g. n1=n2=200, M1=M2=1000, f1=.2, neglect other terms)

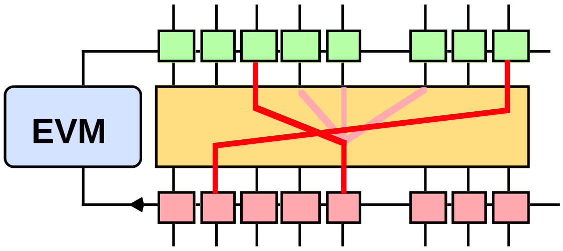
L1	DAQ Level-1 trigger rate	(100 kHz)
N	Total number of RUFU units	(512)
l1	FU trigger rate	(200 Hz)
Pb	Bandwidth (bits/s) required at each port	
n1	Number of messages to be exchanged for HLT level 2	
M1	Average size (bits) of message for HLT-2	
f1	Fraction of events passing HLT-2	
n2,M2,f2	The same for HLT-3 etc.	

Event builder protocols



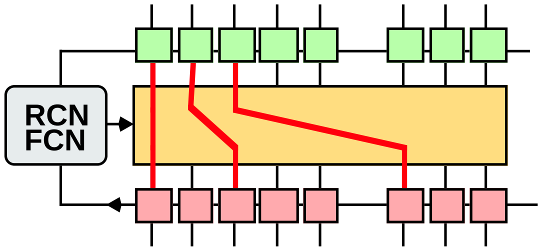
PACKET SWITCHING(PUSH)

- E.g. ATM and FCS (high classes)
- Any size event fragments
- Adapter layer is complex
- Large system performance ? Latency?
- Data congestion controls ?
(traffic shaping, back pressure.)



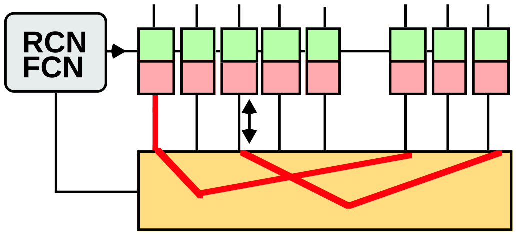
CIRCUIT SWITCHING (PUSH)

- E.g. FCS, HIPPI, custom
- Large size fragments (multi-events)
- Cross bar like switches.
Open connection and send data
- Channel auto selection or from central system (e.g. barrel switch)
- Large system cross bar ?



BARREL SHIFTER (Path manager. PUSH)

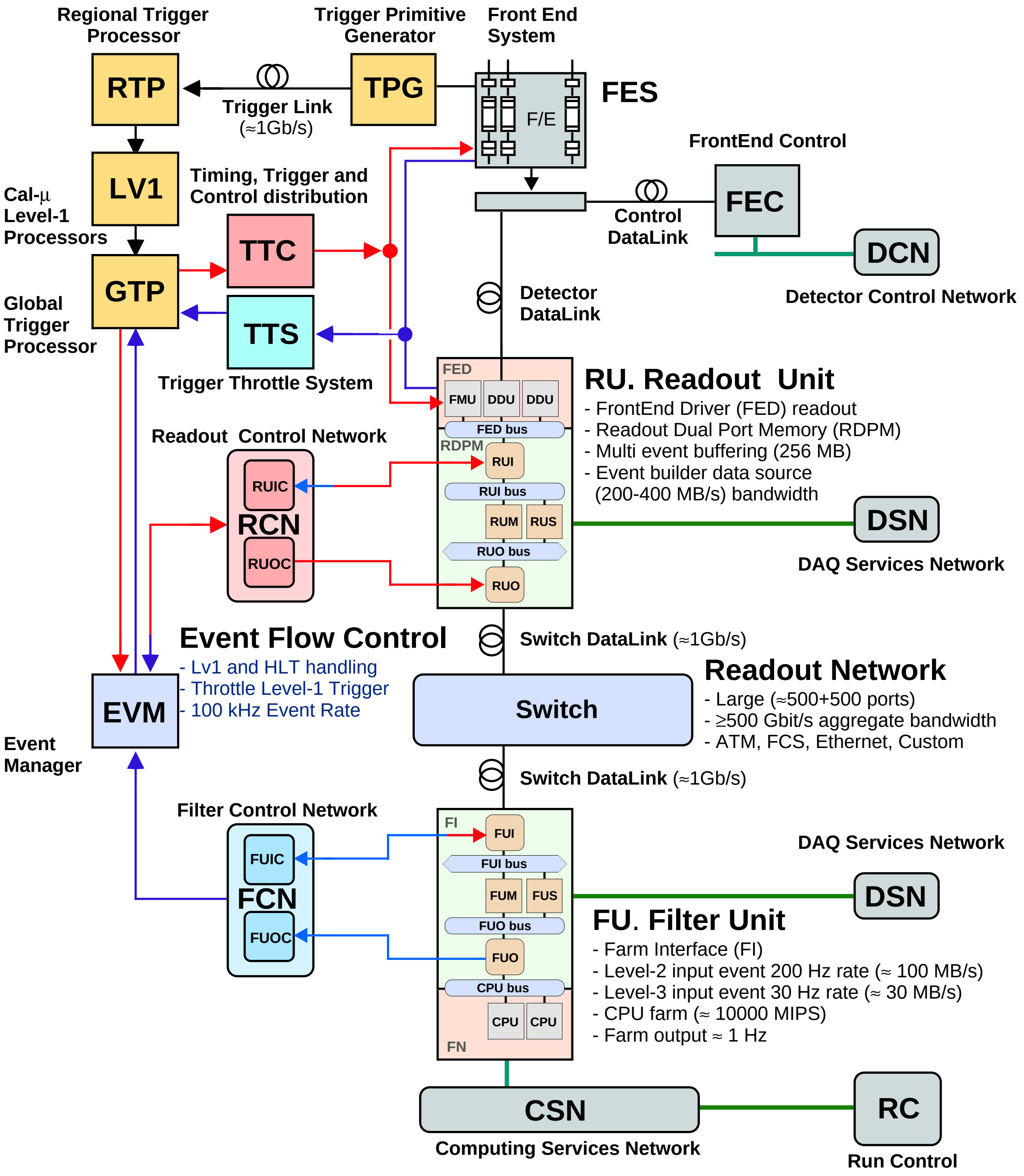
- External node switch allocation
- Custom system using industry nodes
EVM reduced to control network functions



HPCC High Performance Computing and Communication (PULL)

- e.g. Multi Port Memory architectures, SCI, Myrinet.
- RU and FU in the farm node. Event filtering by OO paradigm
- No event flow control. Move data only when needed. EVM reduced to control network functions
- Broadcasting still needed

TriDAS subsystems



Regional Trigger Processor

RTP

Trigger Primitive Generator

TPG

Trigger Link (≈1Gb/s)

Front End System

F/E

FES

FrontEnd Control

FEC

Control DataLink

DCN

Detector Control Network

Cal-µ Level-1 Processors

LV1

Timing, Trigger and Control distribution

TTC

GTP

Global Trigger Processor

TTS

Trigger Throttle System

Detector DataLink

RU. Readout Unit

- FrontEnd Driver (FED) readout
- Readout Dual Port Memory (RDPM)
- Multi event buffering (256 MB)
- Event builder data source (200-400 MB/s) bandwidth

FED, FMU, DDU, DDU, FED bus, RDPM, RUI, RUI bus, RUM, RUS, RUO bus, RUO

Readout Control Network

RUIC, RCN, RUOC

DSN

DAQ Services Network

Event Flow Control

- Lv1 and HLT handling
- Throttle Level-1 Trigger
- 100 kHz Event Rate

Switch DataLink (≈1Gb/s)

Readout Network

- Large (≈500+500 ports)
- ≥500 Gbit/s aggregate bandwidth
- ATM, FCS, Ethernet, Custom

EVM

Event Manager

Switch

Switch DataLink (≈1Gb/s)

Filter Control Network

FUIC, FCN, FUOC

FI, FUI, FUI bus, FUM, FUS, FUO bus, FUO, CPU bus, CPU, CPU, FN

DAQ Services Network

DSN

FU. Filter Unit

- Farm Interface (FI)
- Level-2 input event 200 Hz rate (≈ 100 MB/s)
- Level-3 input event 30 Hz rate (≈ 30 MB/s)
- CPU farm (≈ 10000 MIPS)
- Farm output ≈ 1 Hz

CSN

Computing Services Network

RC

Run Control

FES	**Front End System**
TPG	Trigger Primitive Generator
TDL	Trigger Data Link
RTP	Regional Trigger Processor
GTP	Global Trigger Processor
TTC	Timing, Trigger & Control
TTS	Trigger Throttle System
FEC	**FrontEnd Control**
CDL	Control Data Link
DCN	Detector Control Network
RC	**Run Control**
CSN	Computing&Services Network
DSN	DAQ Services Network
RU	**Readout Unit**
FED	FrontEnd Driver
DDL	Detector Data Link
FMU	Fast Monitoring Unit
DDU	Detector Dependent Unit
RDPM	Readout Dual Port Memory
RUI	Readout Unit Input
RUM	Readout Unit Memory
RUO	Readout Unit Output
RUS	RU Supervisor
RUIC	RU Input Controller
RUOC	RU Output Controller
FU	**Filter Unit**
FI	Filter interface
FN	Filter Node
FUI	Filter Unit Input
FUM	Filter Unit Memory
FUO	Filter Unit Output
FUS	FU Supervisor
FUIC	FU Input Controller
FUOC	FU Output Controller
EVM	**Event Manager**
SDL	Switch Data Link
RCN	Readout Control Network
FCN	Farm Control Network

Towards TDR and plan of work

CMS trigger and data acquisition summary

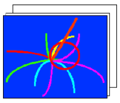
Technology trends

TriDAS schedule

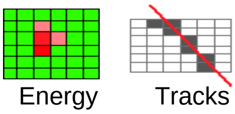
1998-2001 plan of work

CMS trigger and data acquisition summary

COMMUNICATION



40 MHz
COLLISION RATE



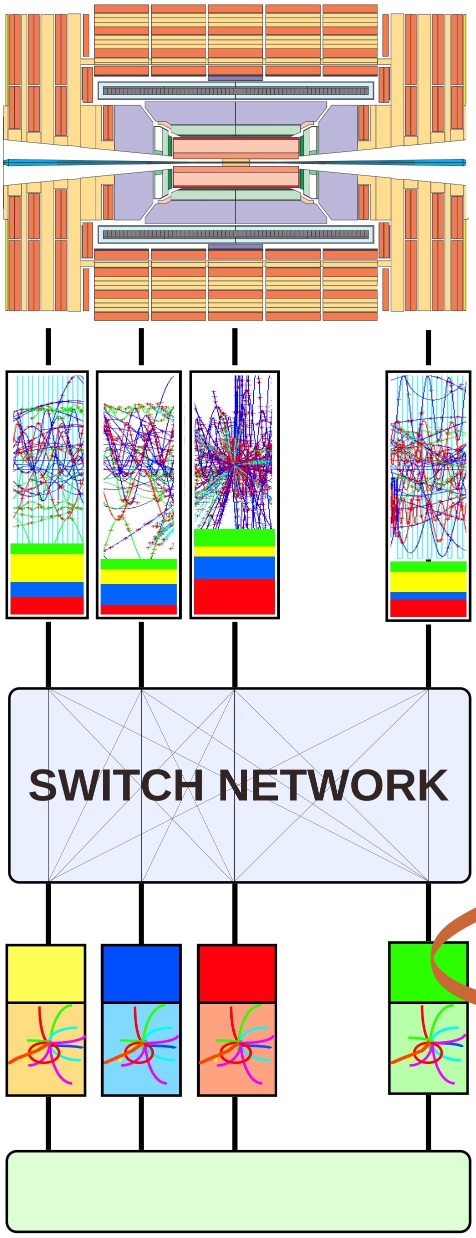
100 kHz
LEVEL-1 TRIGGER

1 Terabit/s
(50000 DATA CHANNELS)

500 Gigabit/s

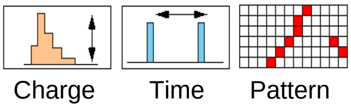
100 Hz
FILTERED EVENT

Gigabit/s
SERVICE LAN



PROCESSING

16 Million channels
3 Gigacell buffers



1 Megabyte EVENT DATA

200 Gigabyte BUFFERS
500 Readout memories

EVENT BUILDER. A large switching network (512+512 ports) with a total throughput of approximately 500 Gbit/s forms the interconnection between the sources (Readout Dual Port Memory) and the destinations (switch to Farm Interface). The Event Manager collects the status and request of event filters and distributes event building commands (read/clear) to RDPMs

5 TeraIPS

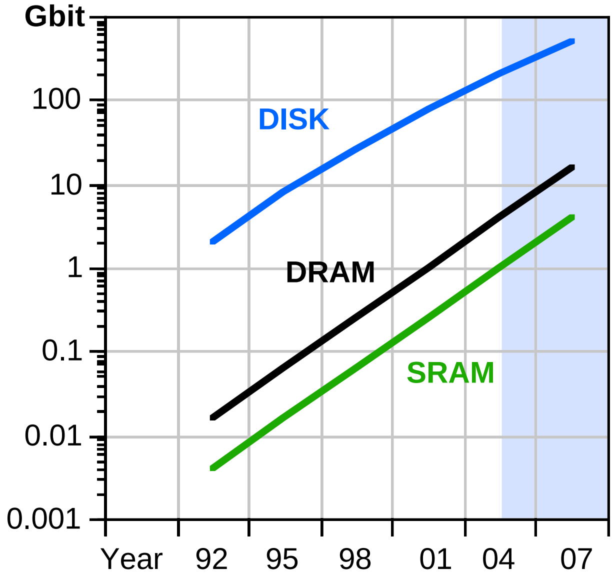
EVENT FILTER. It consists of a set of high performance commercial processors organized into many farms convenient for on-line and off-line applications. The farm architecture is such that a single CPU processes one event

Petabyte ARCHIVE

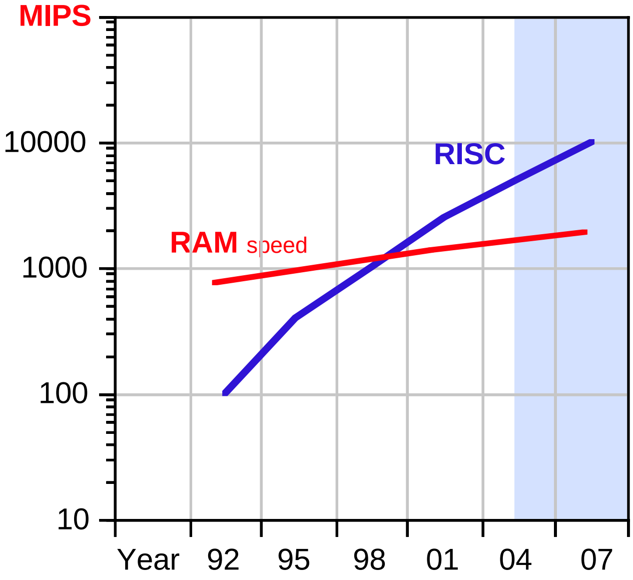
At the LHC, the proton beams cross each other 40,000,000 times each second. At the highest LHC beam intensities, there will be roughly 25 proton-proton collisions for each crossing. Recording all the information from these collisions in the CMS experiment, requires, for every second of operation, the equivalent of 10,000 Encyclopaedia Britannica.
The task of the Trigger and Data Acquisition System is to select, out of these millions of events, the most interesting 100 or so per second, and then store them for further analysis.
An event has to pass two independent sets of tests, or Trigger Levels, in order to pass the TriDAS examination. The tests range from simple and of short duration (Level-1) to sophisticated ones requiring significantly more time to run (High Levels 2, 3, ...)

Technology trends

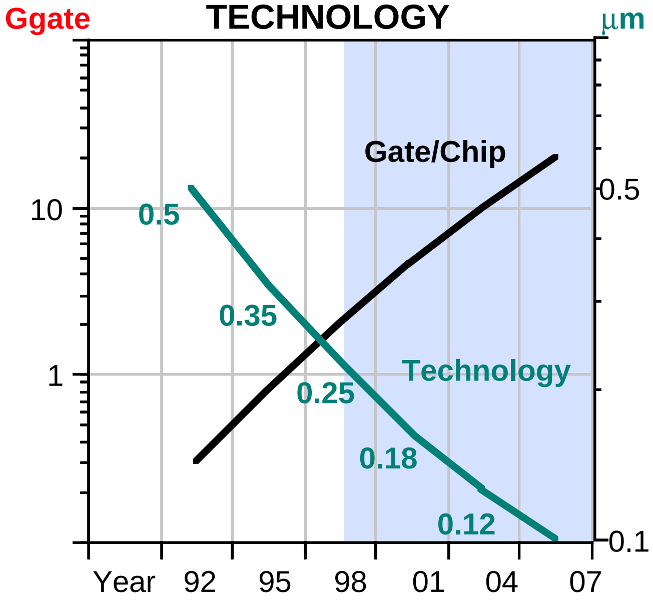
MEMORY DENSITY



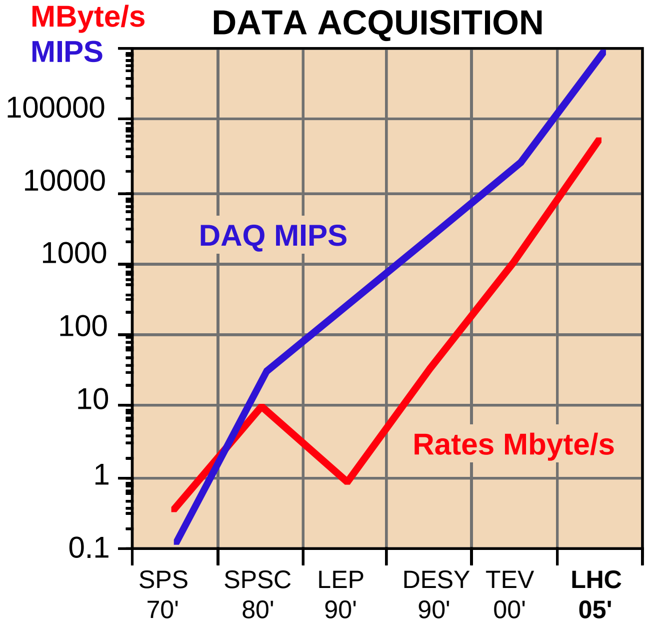
CPU POWER



TECHNOLOGY



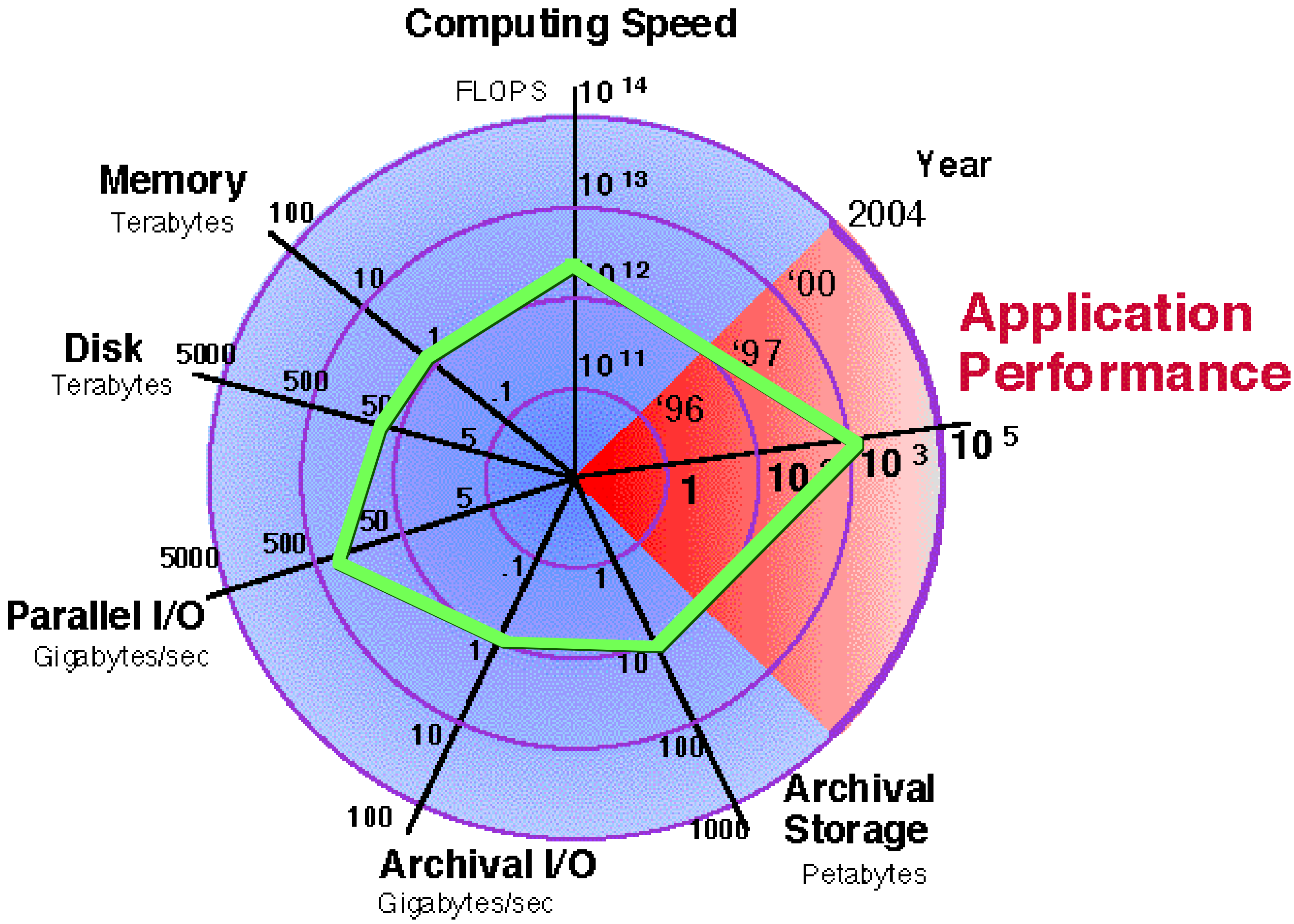
DATA ACQUISITION



- The CPU processing power increases by a factor 10 every 5 years
- Memory density increases by a factor 4 every two years
- The 90's is the data communication decade

ASCI(*) trends

(*) Accelerated Strategic Computing Initiative. ASCI 97



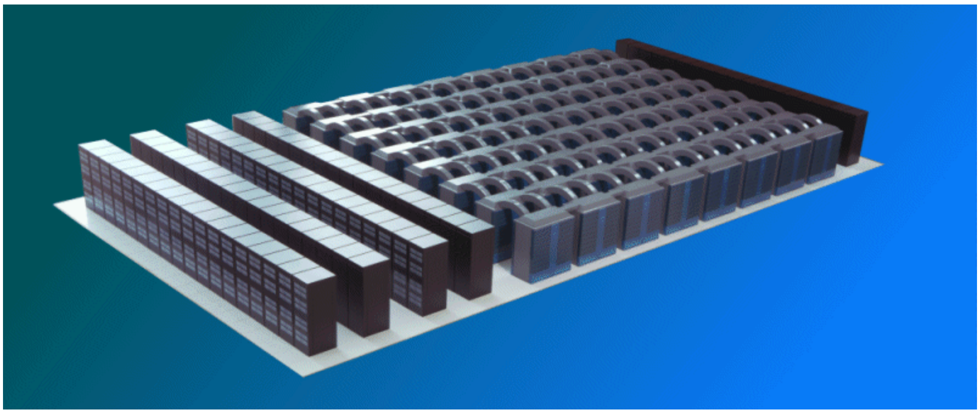
Estimated Computational Resource Scaling Factors

1	FLOP/s	Peak Compute
1	Byte/FLOP/s	Memory
50	Byte/FLOP/s	Disk
0.05	Byte/FLOPS/s	Peak Disk Parallel I/O
0.001	Byte/FLOP/s Peak	Parallel Archive I/O
10000	Byte/FLOP/s	Archive

The computing power, the data storage and the channel bandwidth needed for the CMS data acquisition are already available today in advanced supercomputer structures.
We expect these performances will be available by commodities in the years 2000. Switching network bandwidth (not MPP systems) trends are less clean.

ASCI PathForward Program Overview

ASCI (Accelerated Strategic Computing Initiative) home page URL
http://www.llnl.gov/asci/



The purpose of the Path Forward Project's development and engineering alliances is to ensure the availability of the essential integrating and scaling technologies required to create a well-balanced, reliable and production capable computing environment **providing large scale, scientific compute capability from commodity building blocks, at processing levels of 10-to-30 TFLOPS in the late 1999 to 2001 timeframe and 100 TFLOPS in the 2004 timeframe.** This capability is needed to perform large scale numerical simulations in support of stockpile stewardship. At this time, we are particularly interested in industrial alliances with U.S. companies which will impact 10-to-30 TFLOPS systems. Subsequent PathForward Phases will address 100 TFLOPS systems. Further out in time, an even higher computational capability beyond the petaFLOPS range is needed. However, this is not addressed by the Path Forward Project. **The technologies to be developed under the PathForward Program are based on an underlying approach in which the requisite computational resources are obtained by the aggregation of multiple commodity compute nodes**. The purpose of the PathFoward Project is the development of the unavailable, enabling non-mass market hardware and software technologies that will permit the creation of a balanced 100 TFLOPS compute environment in the 2004 timeframe from mostly commodity hardware/software components. This approach maximally leverages the naturally occurring technical progress driven by mass market forces. **We recognize that the size and detailed implementation of commodity nodes, peripherals, and software is determined by market forces, and it is not a goal of the PathForward Program to drive the market sweet spot.** The PathForward Program itself is dedicated exclusively to the development of those hardware and software components which permit the scaling in computational capability by orders of magnitude beyond market-driven commercial systems.

NEWSCOPE

japan 32tera

NEC

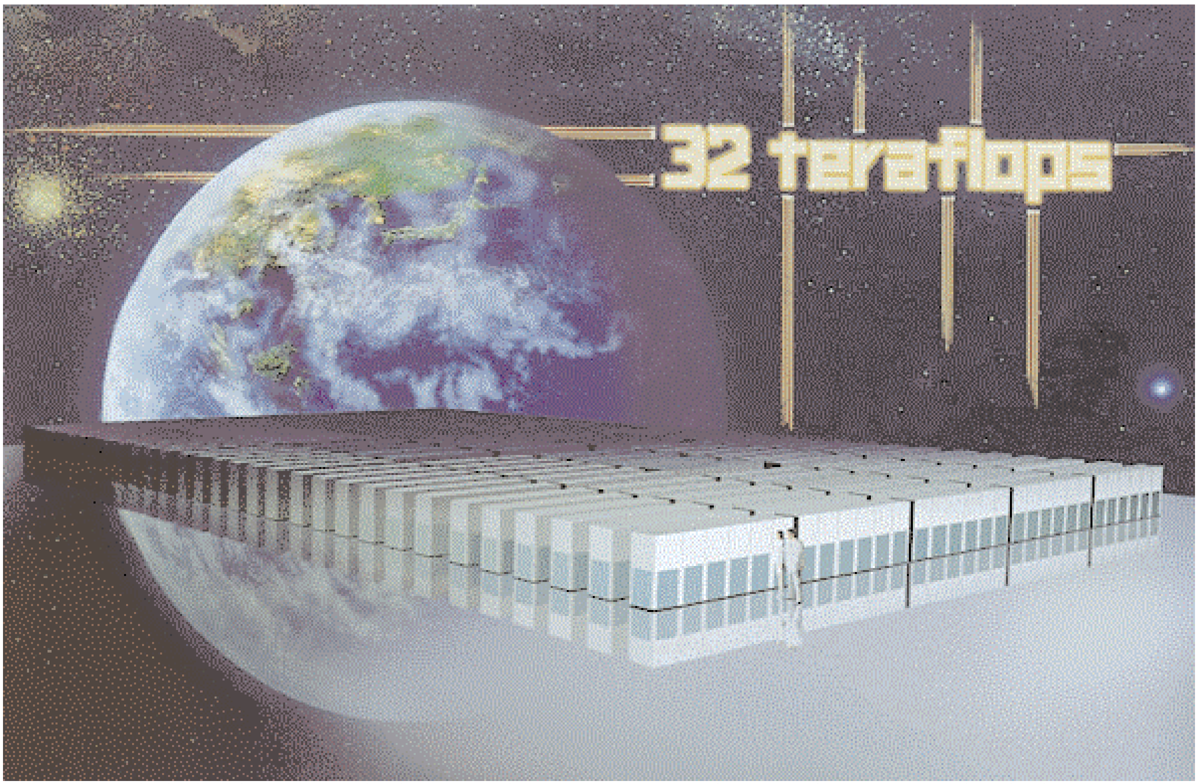


Image of the Earth Simulator (courtesy of National Space Development Agency of Japan/Japan Atomic Energy Research Institute)

MORE INFORMATION
BACK to NEWSCOPE 168 Menu

ULTRA COMPUTER
WILL CREATE 'VIRTUAL EARTH'

Environmental problems such as global warming and atmospheric/marine pollution are now serious concerns of everyone. To achieve more accurate analysis (which can serve as the basis for effective solutions), scientists need a new breed of computers powerful enough for high-resolution modeling of a 'virtual earth'.

Japan's Science and Technology Agency has chosen NEC to create the basic design for an ultra computer capable of simulating advanced, combined models. The ultra computer will play a key role in the Agency's Earth Simulator Program.

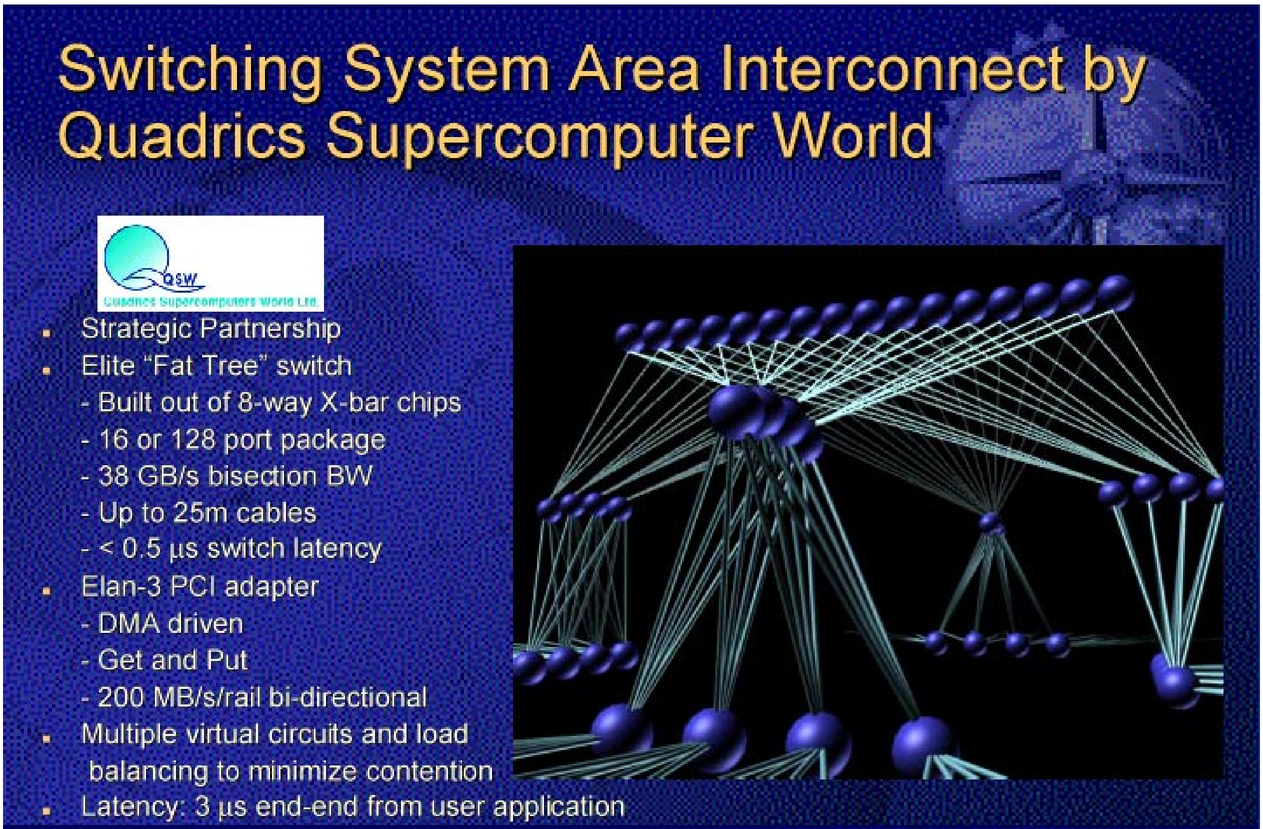
One goal of the Earth Simulator Program is to provide more accurate forecasts of climate change on a global scale. By utilizing data from earth observation satellites, the Simulator will also provide more precise predictions of planetary phenomena, including global warming, atmospheric/marine pollution and El Nino weather.

The ultra computer will be approximately one thousand times faster than the supercomputers now used in meteorological and environmental fields. The peak performance target for the Earth Simulator is over 32 teraflops (32 trillion floating point operations per second). Main memory capacity will be over four terabytes. Designers will attain this unprecedented speed by combining thousands of vector-type CPUs in parallel.

The schedule calls for development of the ultra computer to be completed in the year 2002.

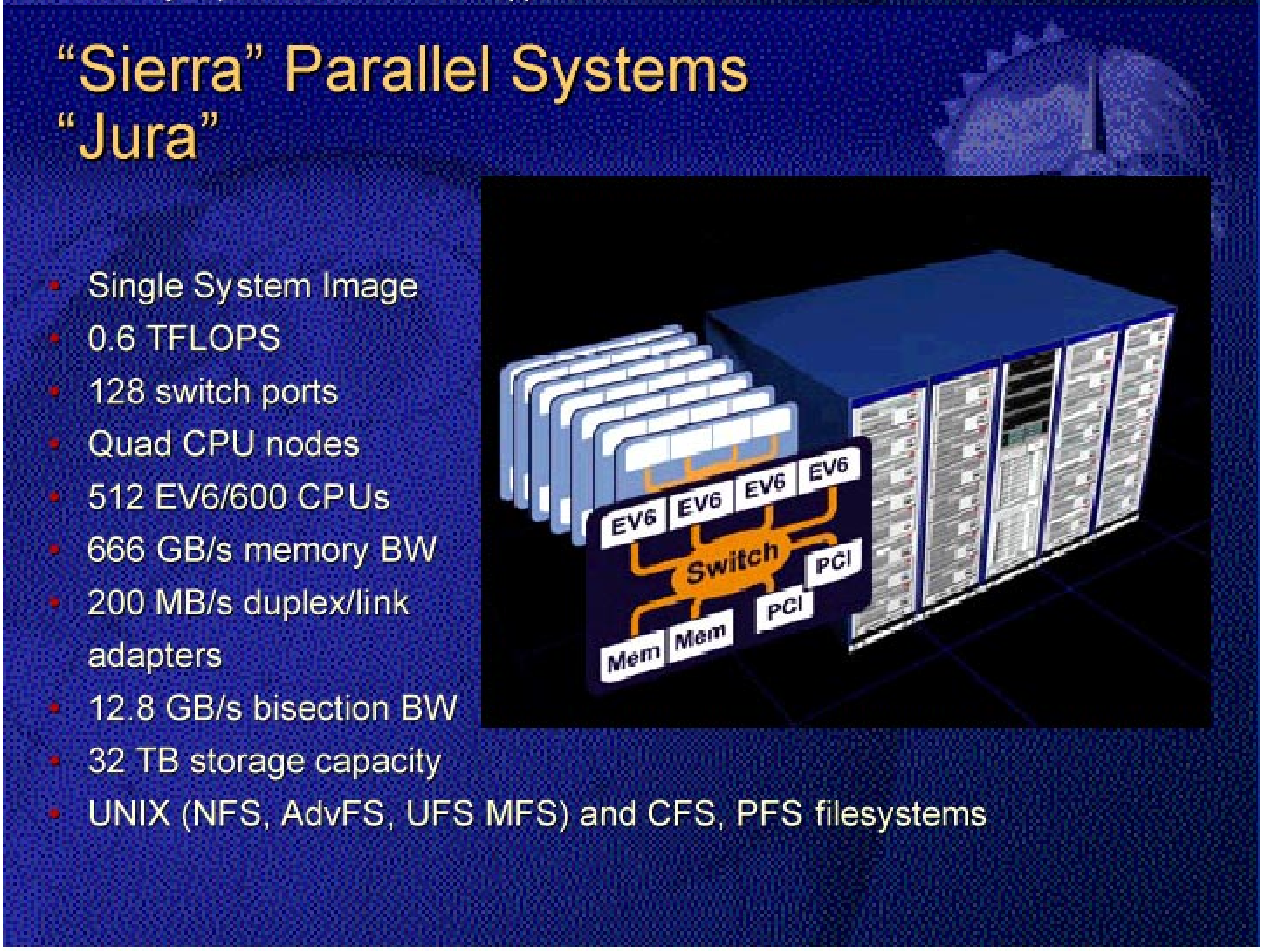
All specifications are subject to change without notice.
All trademarks are the property of their respective holders.

CompaQ HPCC



Switching System Area Interconnect by Quadrics Supercomputer World

- Strategic Partnership
- Elite “Fat Tree” switch
  - Built out of 8-way X-bar chips
  - 16 or 128 port package
  - 38 GB/s bisection BW
  - Up to 25m cables
  - < 0.5 μs switch latency
- Elan-3 PCI adapter
  - DMA driven
  - Get and Put
  - 200 MB/s/rail bi-directional
- Multiple virtual circuits and load balancing to minimize contention
- Latency: 3 μs end-end from user application

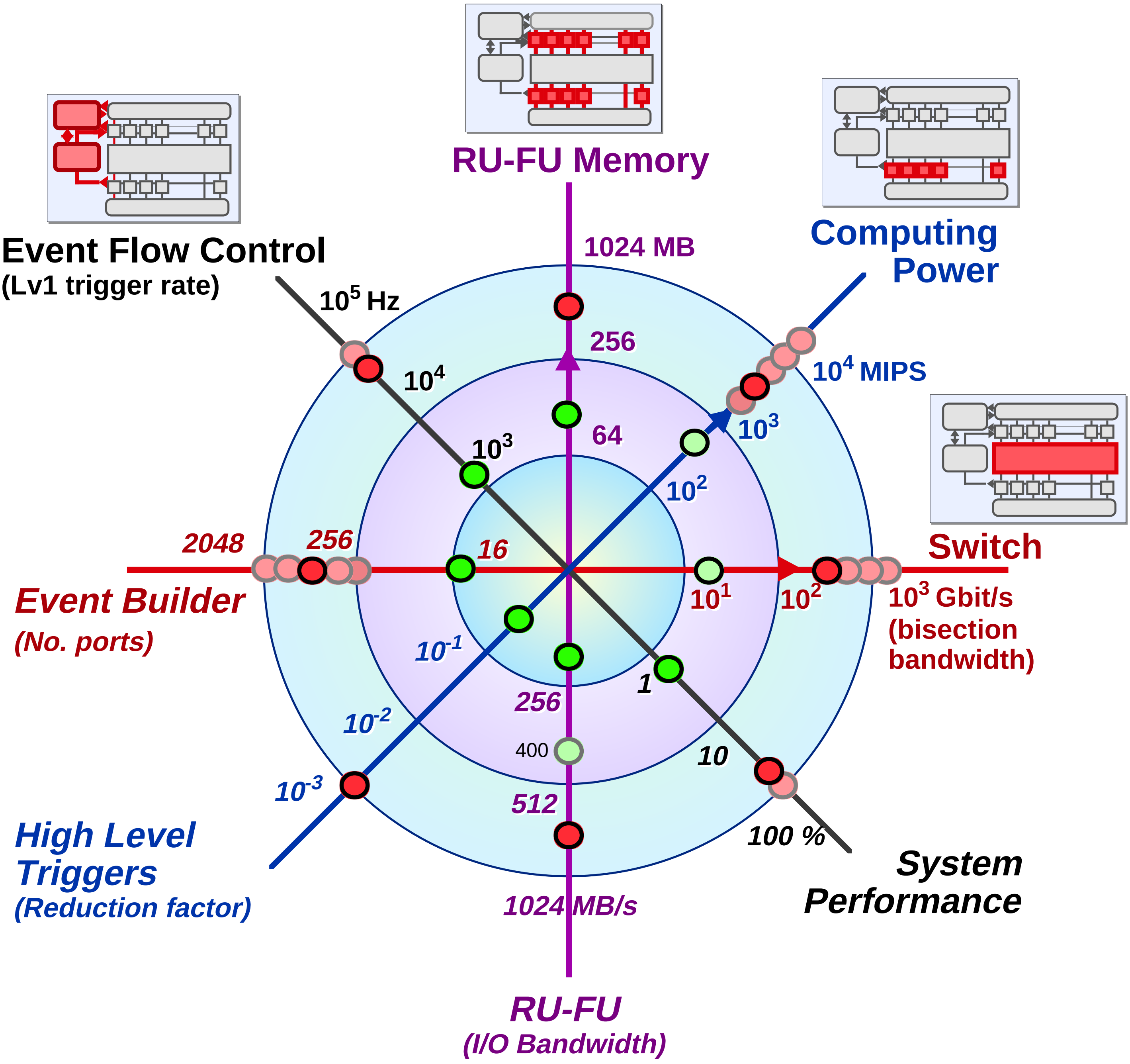


“Sierra” Parallel Systems “Jura”

- Single System Image
- 0.6 TFLOPS
- 128 switch ports
- Quad CPU nodes
- 512 EV6/600 CPUs
- 666 GB/s memory BW
- 200 MB/s duplex/link adapters
- 12.8 GB/s bisection BW
- 32 TB storage capacity
- UNIX (NFS, AdvFS, UFS MFS) and CFS, PFS filesystems

EV6
EV6
EV6
EV6
Switch
PCI
PCI
Mem
Mem

DAQ parameters & prototypes status



TriDAS schedule

1996-2001   Technical design

Identify functions and subsystems by prototyping
Select technologies and options by integration of Labs test benches and demonstrators in test beams

2001-2002   Demonstrator

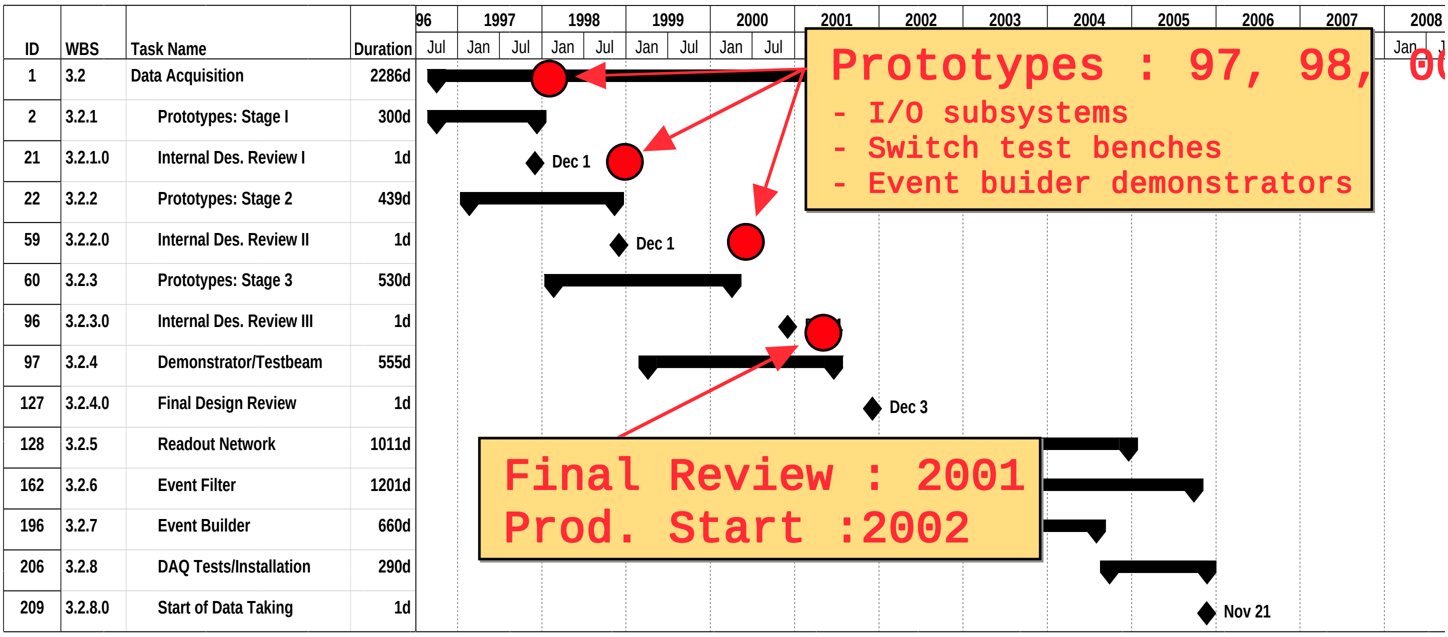
32 X 32(*) Event Builder; Full DAQ prototype in test beam

2001-2004   Construction

System engineering, production, tests, purchasing, installat
and detector subsystems integration.
On-line software development.
Documentation, training

2005- ...   Operation

Start data taking. Define upgrade programme



ID	WBS	Task Name	Duration
1	3.2	Data Acquisition	2286d
2	3.2.1	Prototypes: Stage I	300d
21	3.2.1.0	Internal Des. Review I	1d
22	3.2.2	Prototypes: Stage 2	439d
59	3.2.2.0	Internal Des. Review II	1d
60	3.2.3	Prototypes: Stage 3	530d
96	3.2.3.0	Internal Des. Review III	1d
97	3.2.4	Demonstrator/Testbeam	555d
127	3.2.4.0	Final Design Review	1d
128	3.2.5	Readout Network	1011d
162	3.2.6	Event Filter	1201d
196	3.2.7	Event Builder	660d
206	3.2.8	DAQ Tests/Installation	290d
209	3.2.8.0	Start of Data Taking	1d

Prototypes : 97, 98, 0
- I/O subsystems
- Switch test benches
- Event buider demonstrators

Final Review : 2001
Prod. Start :2002

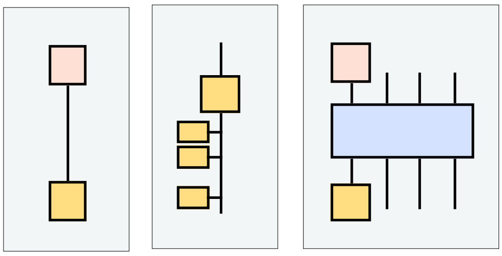
The prototyping programme covers the most critical points and design options of the CMS trigger and dat acquisition system. It will include sub-systems demonstrators and test benches for technology evaluation

Towards TDR

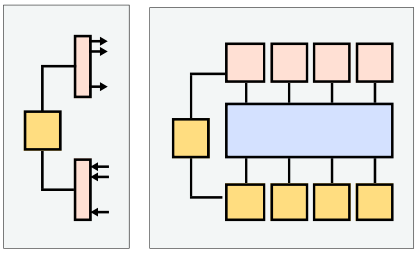
MEDIUM TERM PROGRAM (--> 2001)

- Evaluation and choice of technologies
- Basic DAQ subsystems development
- Next generation test beam data acquisition
- Technical Design Report

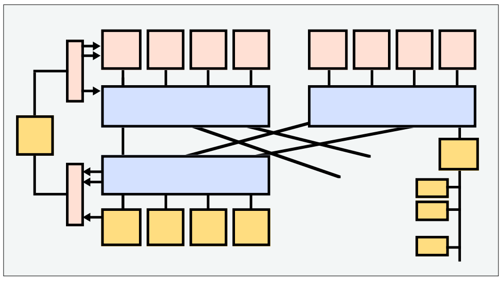
PROTOTYPE PLAN:



1996-1997. Subsystem prototyping of data handling architectures, interfaces and processing elements



1998-1999. Prototype systems integration (DAQ demonstrators). Functionality studies. Setting up of test benches for technology evaluation and studies of protocols and structures. Extrapolate large system behavior by simulation. Transfer H/S prototypes into test beam DAQ systems



2000-2001. Full DAQ demonstrators
Full data acquisition system scale 1:32.
Design options and technologies choice
Technical Design Report **TDR**

2002-2005 System engineering, production, tests, installation and detector subsystems integration. On-line software development.

1998-2001 plan of work

Design, simulate and prototype the hardware and software basic components of the CMS baseline data acquisition system (readout unit, filter unit, readout network and event builder). Integrate the above subsystems in test benches (DAQ demonstrators) to evaluate the design options and the technologies in the field of programmable logic systems, bus interfaces, data links and switches, real time operating systems and modern software engineering. Simulate the previous subsystems as single units and as a global integrated DAQ, cross check with the results of laboratory test beds and compute performances of large-scale systems.

The following systems are going to be studied:

- Readout Unit (RU). Software and hardware implementations of RU (and farm interfaces) to be used to identify a modular design with optimized hardware and software partitions.

- Filter Unit (FU). Software framework to study the application of OO methodologies to on-line analysis. To evaluate both mass storage based on OO databases and data communication software technologies to control a large complex of processing elements (Farm controls).

- Generic switch-based event builder. To provide an environment in which we can test: various switching technologies (e.g. ATM Vs Fiber Channel), various switches from different manufacturers (e.g. Ancor Vs Brocade) and various DAQ protocols (e.g. EVM or Farm-based synchronization).

- Integrate all DAQ elements (RU, FU) into a full vertical readout chain to be used in test beams and in laboratory demonstrators

- Simulation of the structure of the front-end readout chain.

Study the design of interface between the detector electronics and DAQ, evaluation of critical data flow parameters (latency, queue depth, back pressure signals, error recovery) for each detector readout.

- High Level Triggers (HLT). Systematic study of the partitioning of detector data into different RU configuration and the corresponding algorithms. Comparison of the background rate rejection and efficiency for physics signals for various algorithms. Optimization of the detector readout partition and evaluation of the reduction of the event rate accepted by the Level-1 trigger.

The final result of the studies during 2000-2001 will be the Data Acquisition Technical Design Report.

Prototypes and milestones

December 98 milestones

RU(FU): 98 milestones 1

RU(FU): 98 milestones 2

DAQ demonstrators 98 milestones

High Level Triggers 98 milestones

DAQ 98 demonstrators summary

December 98 milestones

Readout subsystem prototypes and test bench installations.

RU prototype 1: The RU prototype will provide the full functionality required for the "event_ID" protocol.
- RUM: the functionality will be provided by the reprogrammed RDPM21++ module supplemented with two auxiliary PMC cards to furnish a true dual-ported access to the memory (via dual PCI busses).
- **RUI/RUO:** the functionality will be provided by commercial VME CPUs running VxWorks and supplemented with a PMC card each (to connect to the RDPM21++)
- Software: This is the software that will run under VxWorks in the RUI and RUO CPUs, suitable for use in both the EVB test-bench and in small DAQ systems.

VxRU prototype 1: This prototype of the Readout Unit will be based on a commercial CPU running a (potentially real-time) Operating System. For this milestone, the OS chosen is VxWorks.

Event Builder test-bench: A 4x4 Event Builder based on multiple switch technologies and inputs and outputs (RUs and FUs) resembling the current module developments as closely as possible.
- Switch: both Fibrechannel (ANCOR) and Myrinet switches will be utilized
- EVB prototype simulation: it will provide a functional model of the 4x4 EVB setup.
- RCN/FCN networks: implemented via external Ethernet switches. Both TCP/IP and light protocols will be used.
- Filter Unit: Farm monitoring and control

HLT results on single-e,μ triggers

- First results from simulation and HLT studies; evaluation of rates relative to Level-1 trigger with and without tracking information.

Participation in DAQ-98 prototyping

CERN G. Antchev, E. Cano, S. Cittolin, S. Chatelier, D.Gigi, J. Gutleber, A. Kruse, C. Jacobs, T. Ladzinski, R. Nicolau, L. Orsini, L. Pollet, D. Samyn, P. Sphicas, A. Racz

- Readout/Filter Unit FPGA/Desktop based
- Filter Node software and farm handling
- DAQ demonstrators and simulation
- High level triggers
- TTC system

FNAL E. Barsotti, M. Bowden, V. Odell

- Event flow control, switch system evaluation and design

Legnaro INFN-LNL L. Berti, G. Maron, G. Vedovato

- DAQ demonstrators and DAQ controls

MIT S. Sumorok, S. Tether

- Readout/Filter Unit FPGA/Desktop based
- Event flow control, switch system evaluation and design

RAL B. Haynes, R. Halsall

- Frontend driver development

UCSD J. Branson, M. Mojaver

- Filter Unit developments

UCLA S. Erhan

- DAQ simulation

Helsinki E. Pietarinen

- Level-1 and detector data links evaluation

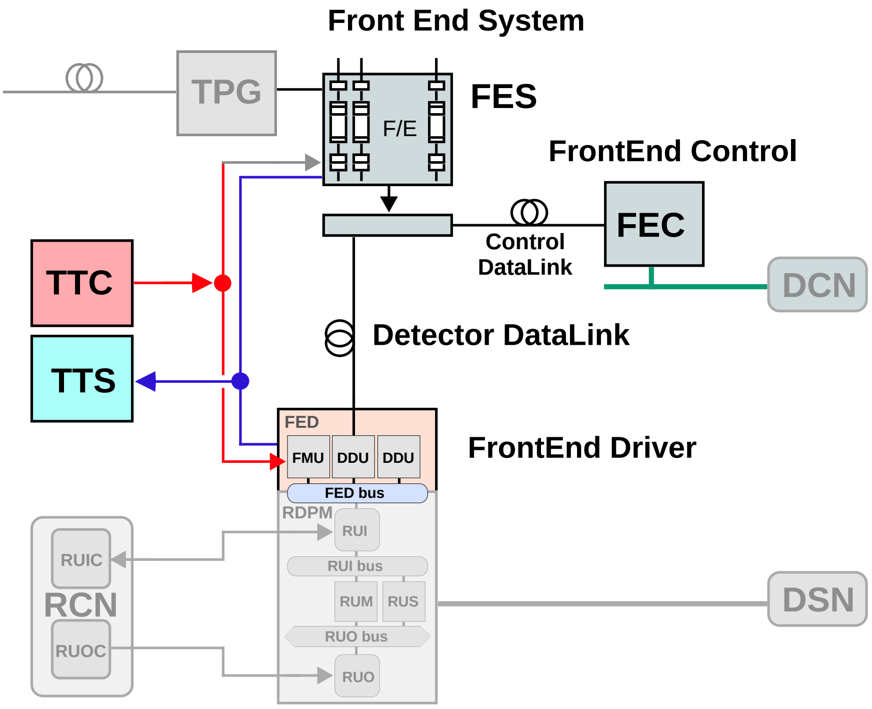
RU(FU): 98 milestones 1

RU prototype 1: The RU prototype will provide the full functionality required for the "event_ID" protocol.
- RUM: the functionality will be provided by the reprogrammed RDPM21++ module supplemented with two auxiliary PMC cards to furnish a true dual-ported access to the memory (via dual PCI busses).
- **RUI/RUO:** the functionality will be provided by commercial VME CPUs running VxWorks and supplemented with a PMC card each (to connect to the RDPM21++)
- Software: This is the software that will run under VxWorks in the RUI and RUO CPUs, suitable for use in both the EVB test-bench and in small DAQ systems.

Report on 98 December milestones

The RDPM-P21 module, which was implemented during 97 using FPGAs, has been reprogrammed for a new control protocol that relies on an event label, the "event_ID", provided by the Event Manager. The old RDPM ports have been upgraded via two auxiliary PMC modules, so that the RDPM appears like a genuine dual-PCI memory with internal buffer (and block) handling. The dual-PCI system has been tested with a PCI data generator and a FiberChannel interface for data output. The event pointers and the memory block allocation tasks have been operated at a Level-1 rate of up to 70 kHz while the DMA was able to operate at 120 MB/s for data blocks of 4 kB and higher. The overheads to fetch the event information and to communicate with the supervisor CPU were 2 and 11 µs respectively. These figures and the communication protocols with the supervisor processors are currently being studied in the context of the future RDPM designs. On the memory front, a new design has introduced two improvements: a factor two in throughput from a 64-bit wide PCI bus and another factor two from a 66 MHz clock speed.
As mentioned above a general purpose PMC-FPGA card has been developed to interface a variety of generic user I/O bus with a PCI port. A special version was used to link the RDPM-P21 output to PCI. This module is also used to prototype digital Detector Dependent Units (DDU) and for the further development and testing of the Readout Units. For more information see
http://cmsdoc.cern.ch/~racz/RUWG/welcome.html

FrontEnd readout



FERU working group

Study of detector frontend logical models

Simulation of full chain of each detector

Identification of critical parameters

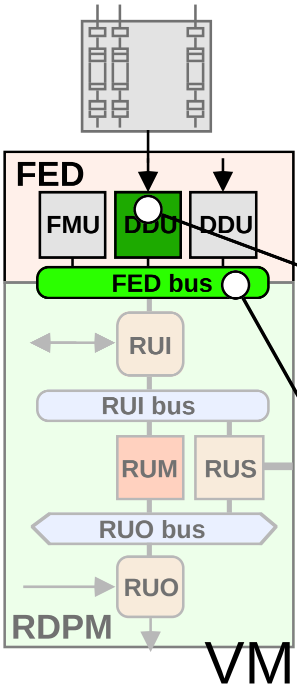
Definition of common interfaces to DAQ

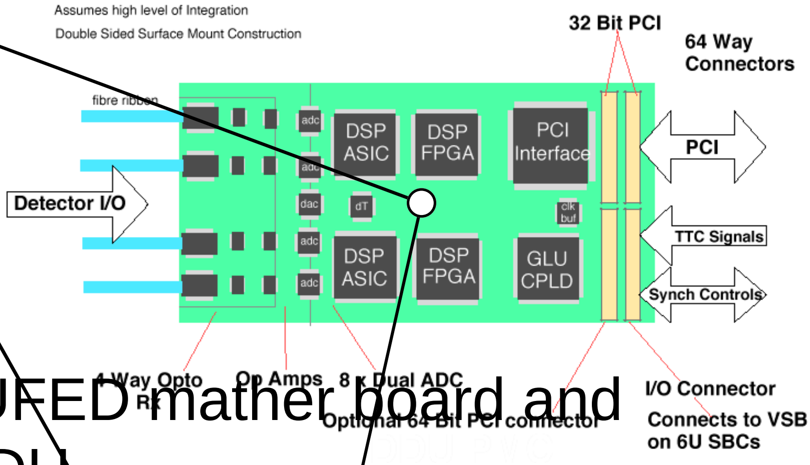
Promotion of common standards and reusable subsytems

Tracker FED prototype (9U VME)

VME 9U mother board (PCI bus support)
Digitizers mezzanines (PMC)
Transition systems:

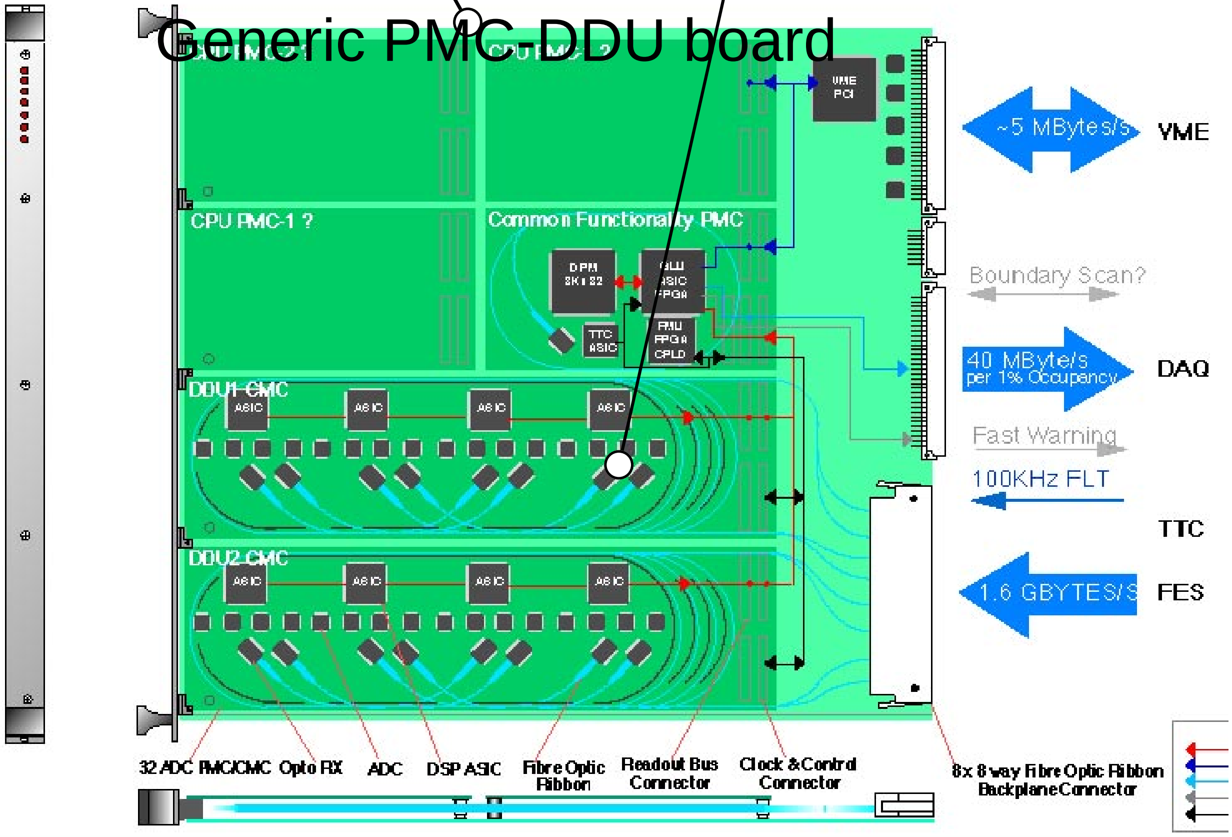
- **Detector links**
- **DAQ (Trigger) links**
- **Fast controls**





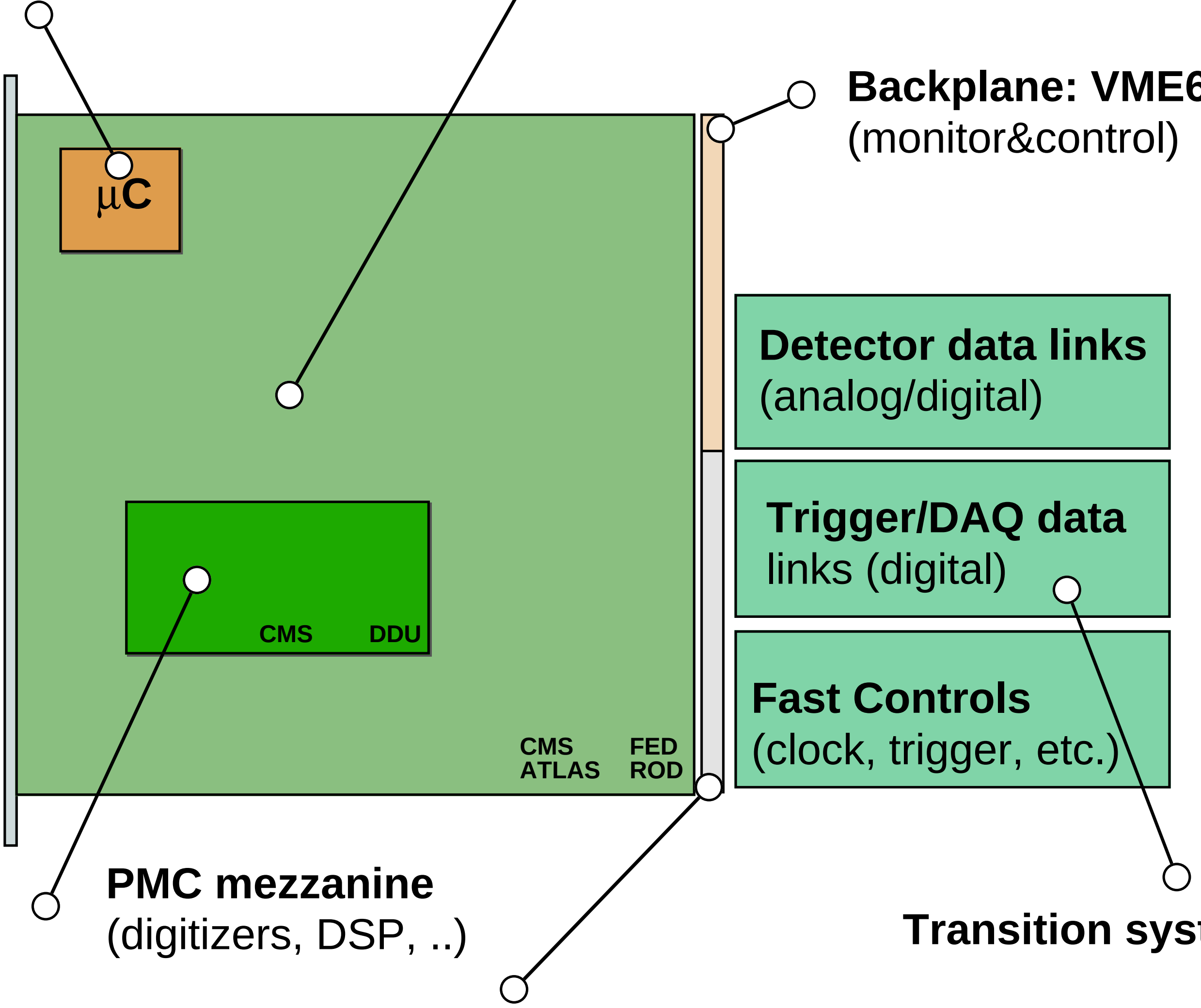
VME 9UFED mather board and PMC-DDU

Generic PMC-DDU board



erford Appleton Laboratory Technology Department System design Group R. Halsall et al

VME 9UFED mather board and PMC-DDU



Embedded microController (ethernet)

VME 9U board. Mainly used as system backplane (e.g. PCI)

Backplane: VME64 (monitor&control)

Detector data links (analog/digital)

Trigger/DAQ data links (digital)

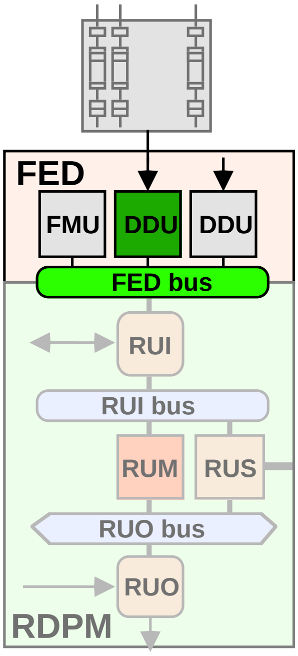
Fast Controls (clock, trigger, etc.)

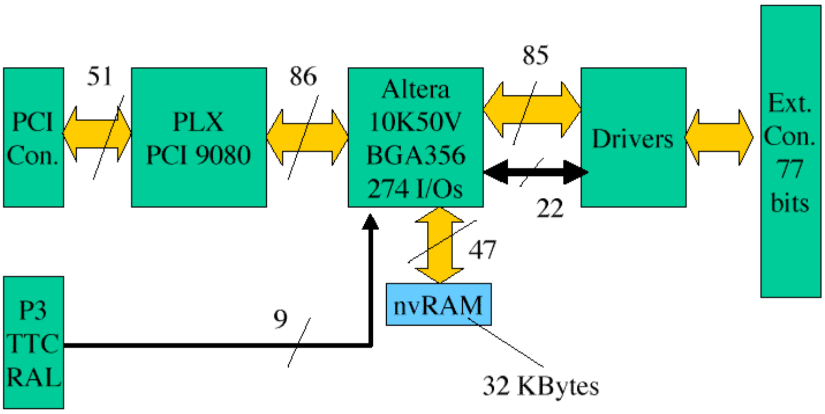
PMC mezzanine (digitizers, DSP, ..)

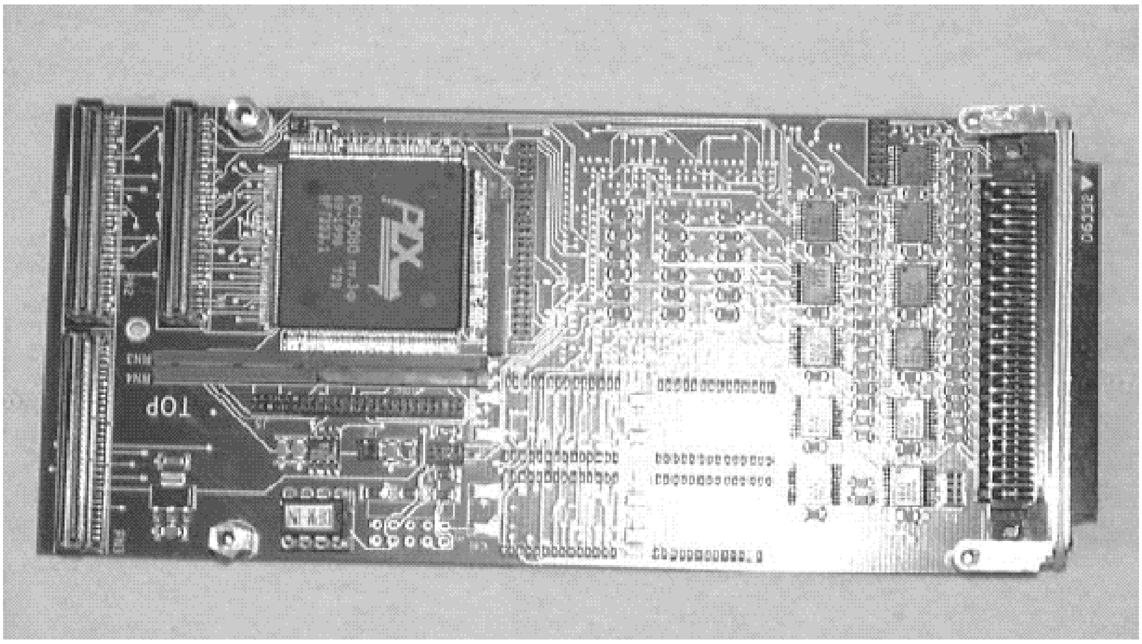
Transition systems

Special backplanes (trigger processors)

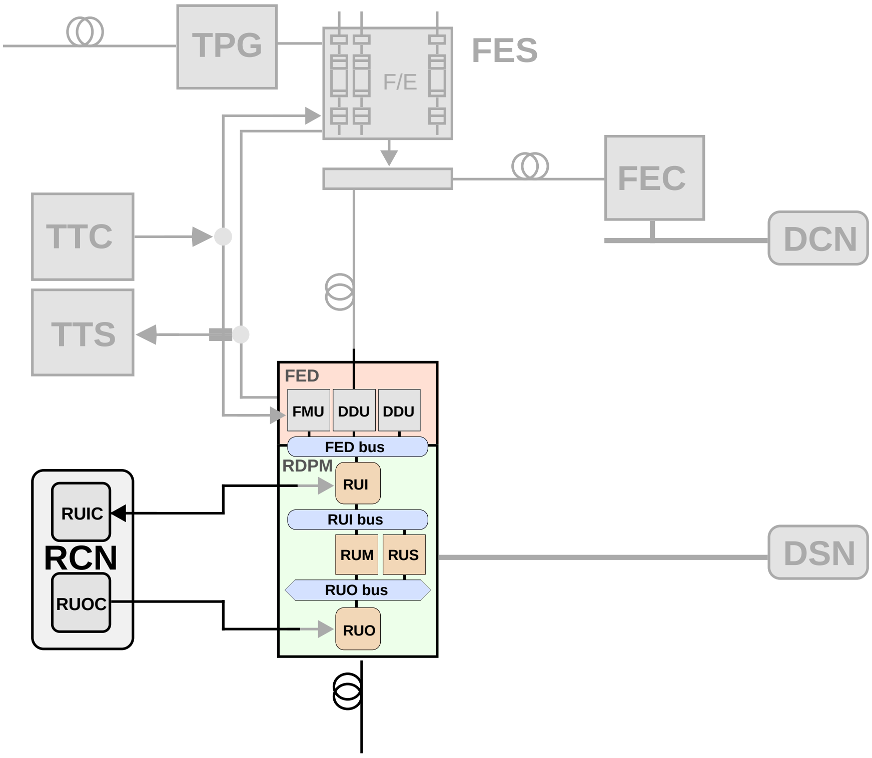
Generic PMC-DDU board







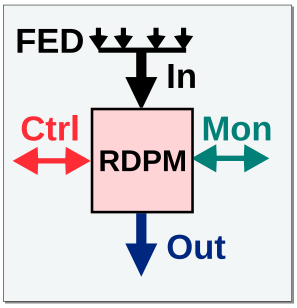
Readout Unit prototypes



RDPM21. FPGA (rewconfigurable computing)

Embedded processor RU and I2O

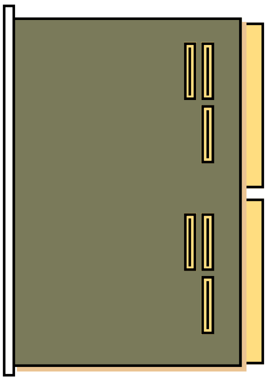
VxWorks readout unit



IN (link/bus)	**400 MB/s**
OUT (link)	**200 MB/s**
Ctrl (network)	**≥10 MB/s**
Mon (network)	**≥10 MB/s**
MEM	**≥256 MB**

PCI form factors

CPU boards (VME/PCI)



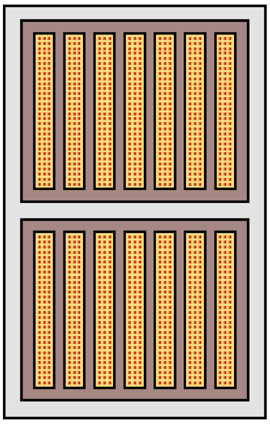
33/66 MHz x 32/64 bit
100/200/400 MB/s

VME and Compact PCI CPU boards with up to 2 slots on one or 2 PCI buses

PMC mezzanine

IO interfaces
Memories

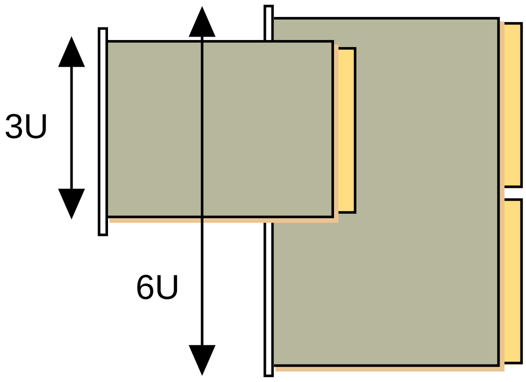
Compact PCI backplanes (3-6 U)



33 MHz x 32/64 bit
100/200 MB/s

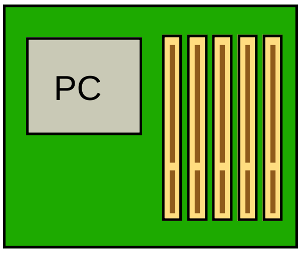
7 Slots backplanes single PCI (3U) double PCI (6U)

Compact PCI modules



CPU, dual CPU, memories and interfaces

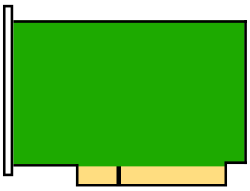
Desktop PCI backplane



33/66 MHz x 32/64 bit
100/200/400 MB/s

Systems with 1, 2 or 3 independent PCI buses

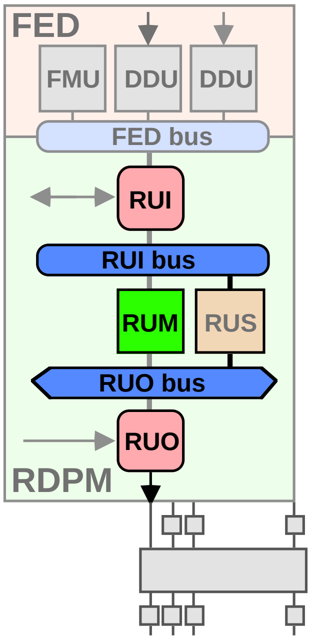
PCI board

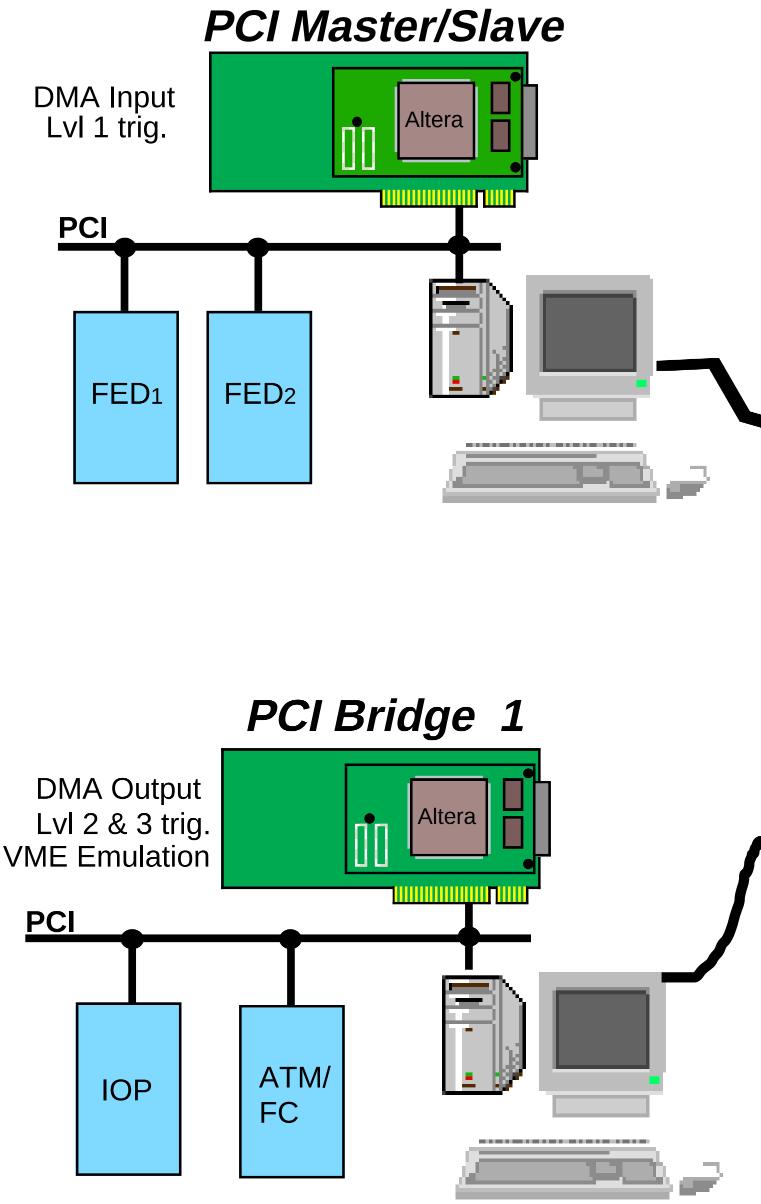


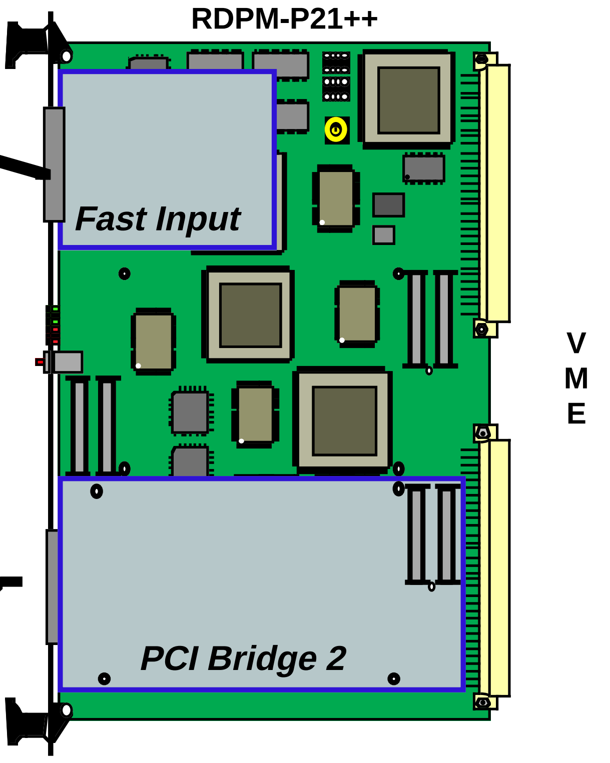
IO interfaces
I2O processors

Readout unit prototypes

1) FPGA design VME (PCI) interface
2) PC desktop
3) Embedded processors (I2O)

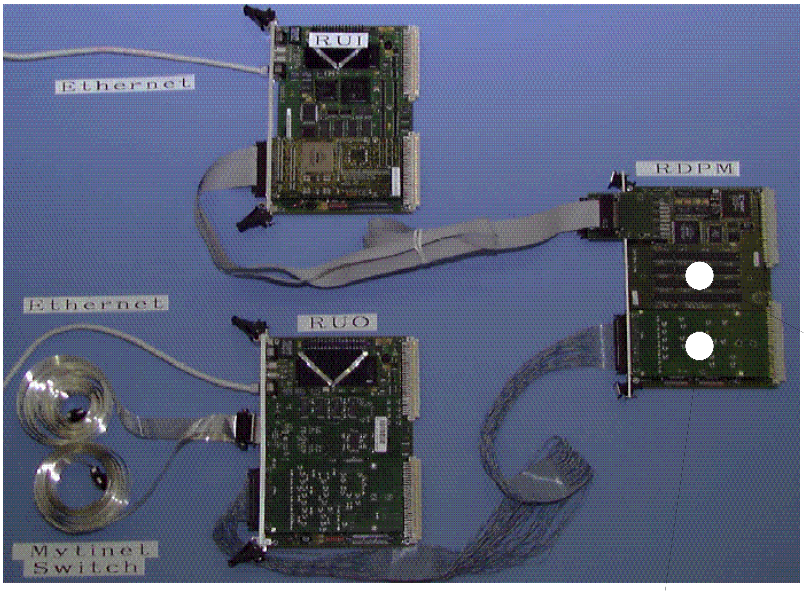




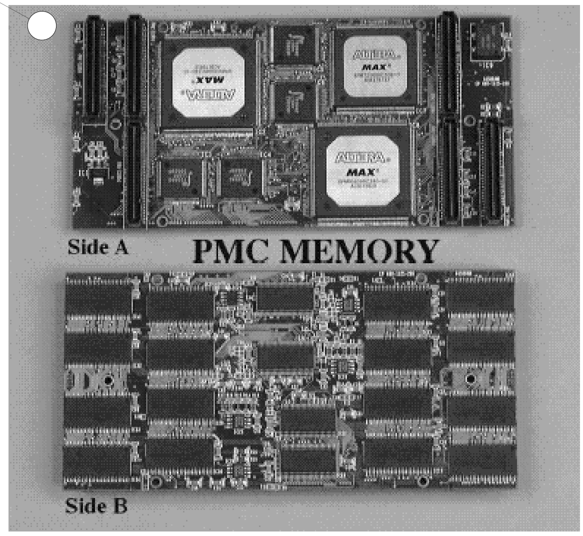


FPGA-PCI-CPU2 prototype

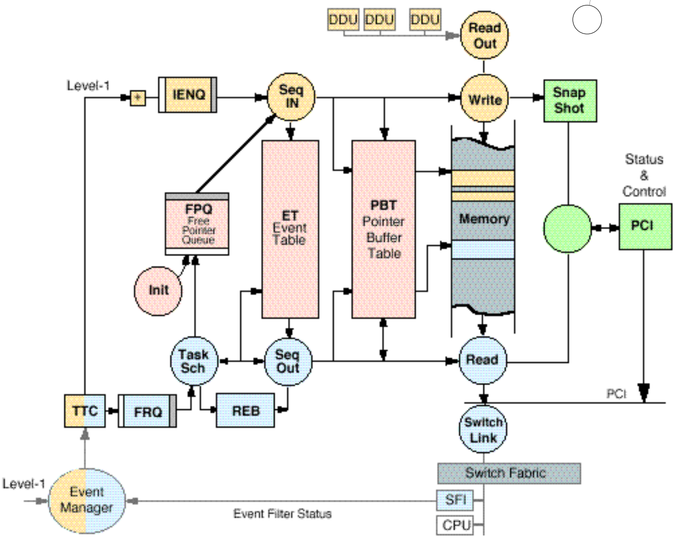
Dual PCI-CPU FPGA RU system



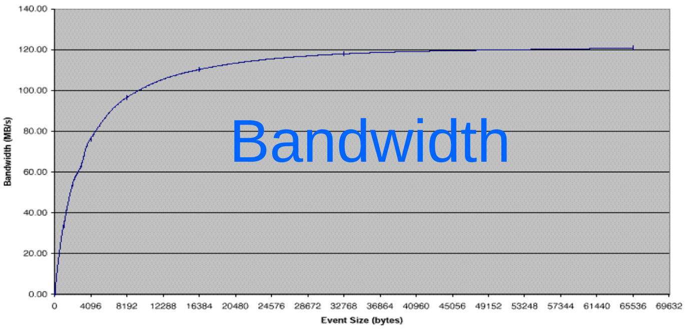
128MB Data memory
200 MB/s 32bit

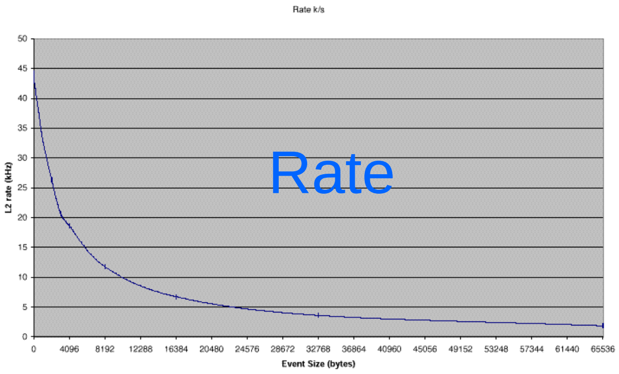


RDPM-P21++ : Functional diagram
Event_ID# Implementation- PCI control



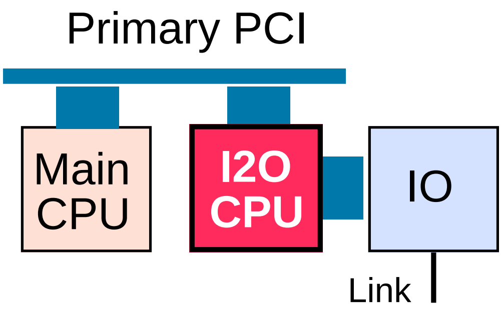
Level-1 and HLT
multi-event handling
logics up to 50 kHz





Input Output Processor IOP (I2O)

Dual PCI bus system with embedded CPU-bridge (e.g. I960r)



CMS IOP:
PCI&VME-PMC
i960r systems

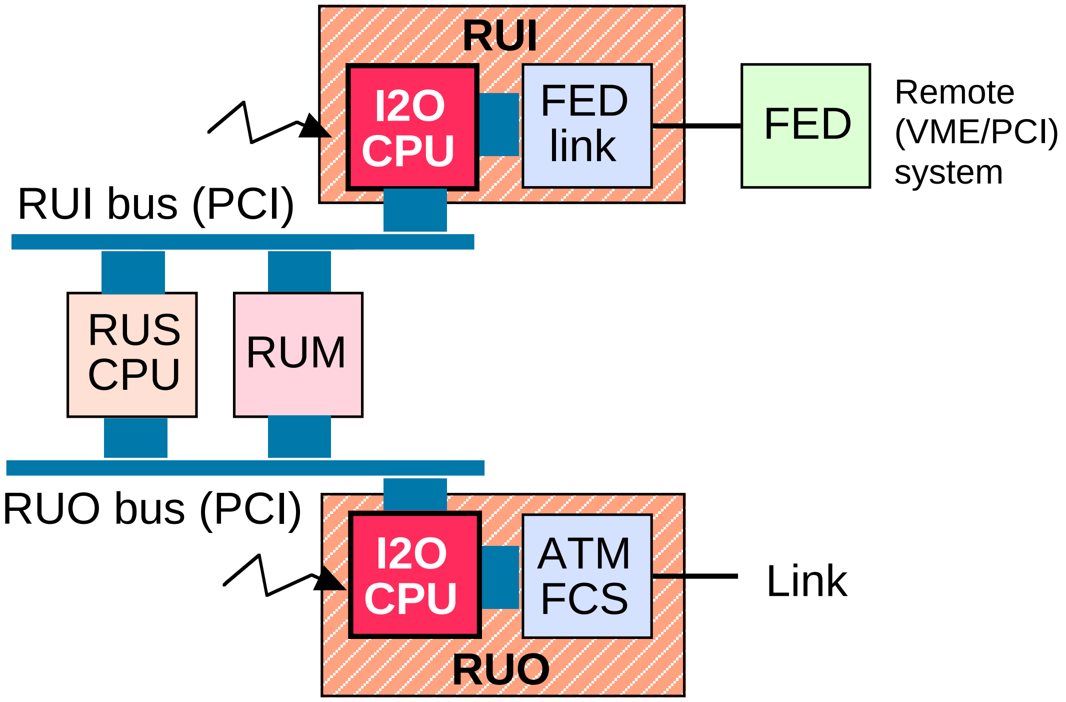
I2O Architecture is intended to solve numerous system performance and systems management problems by:

1. **Off-loading the CPU** (as much as possible) from communications I/O tasks - balancing processing between application intensive (main CPU) and interrupt intensive (I/O processor) operations;

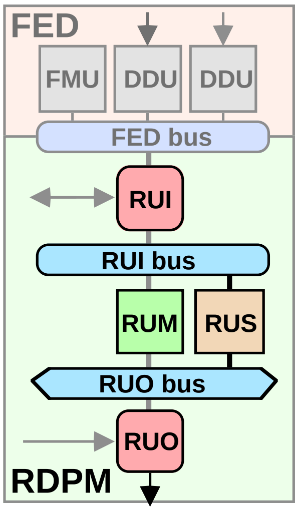
2. **Streamlining the interaction between the I/O** subsystems (communications and storage) - leading to better-optimized performance; and,

3. Providing a **standard interface** for I/O hardware to be managed by the operating environment

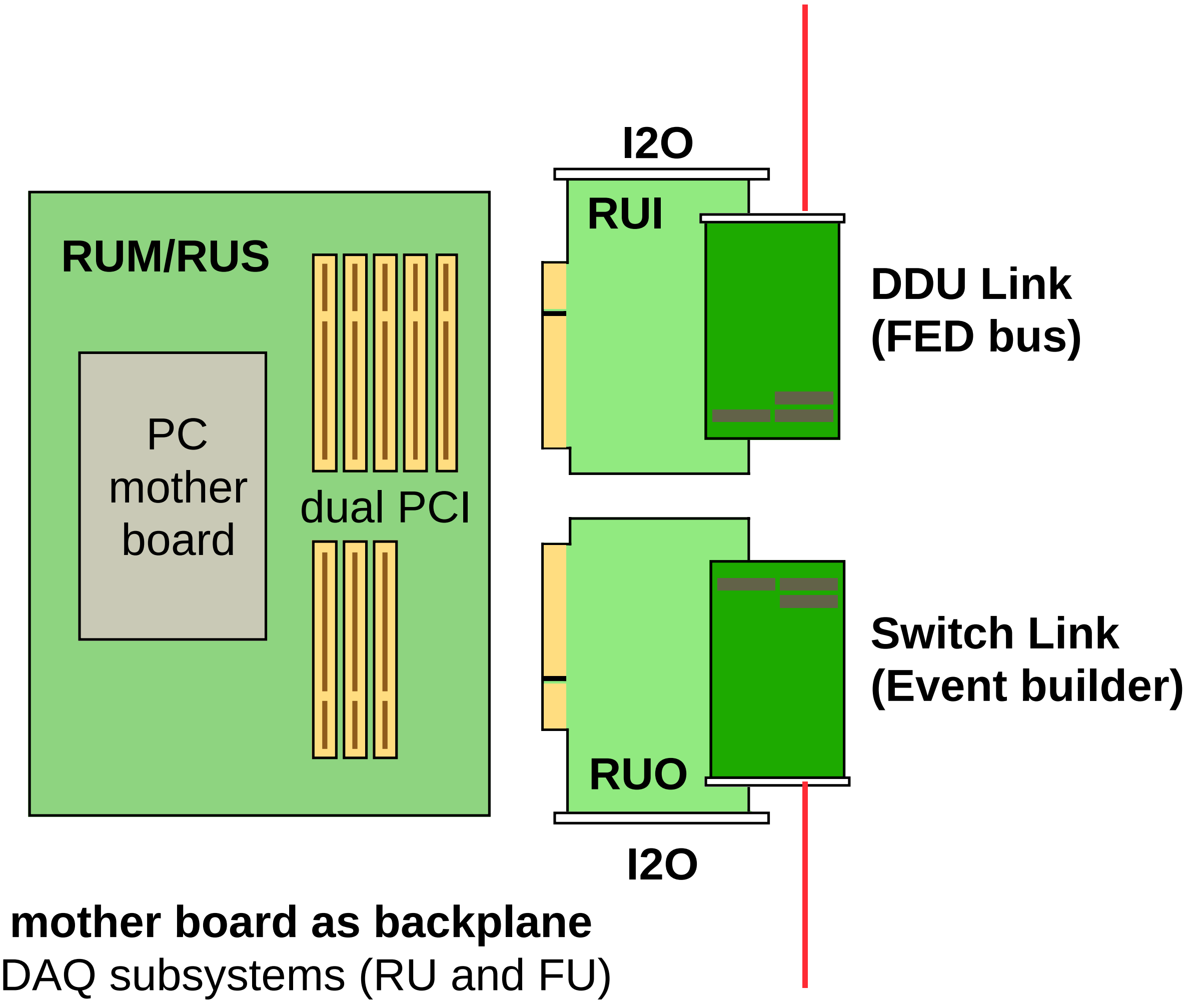
E.g. RUI/RUO systems based on I2O



PCI modularity (DAQ subsystems)



PC based Readout/Filter Units
RU system based on desktop PC. The CPU runs the RDPM event handling logics, the data buffering and the unit monitoring (RUS, RUM). The FED system is readout via a simple PCI-to-PCI connection (PCX) driven by an I2O board (RUI), a second I2O (RUO) drives the output link reading data from the PC memory via PCI. EVM commands (RUIC/RUOC) are distributed via ethernet port in I2Oa(RUI) and I2Ob(RUO).



PC mother board as backplane
for DAQ subsystems (RU and FU)

RU(FU): 98 milestones 2

VxRU prototype 1: This prototype of the Readout Unit will be based on a commercial CPU running a (potentially real-time) Operating System. For this milestone, the OS chosen is VxWorks.

Report on 98 December milestones

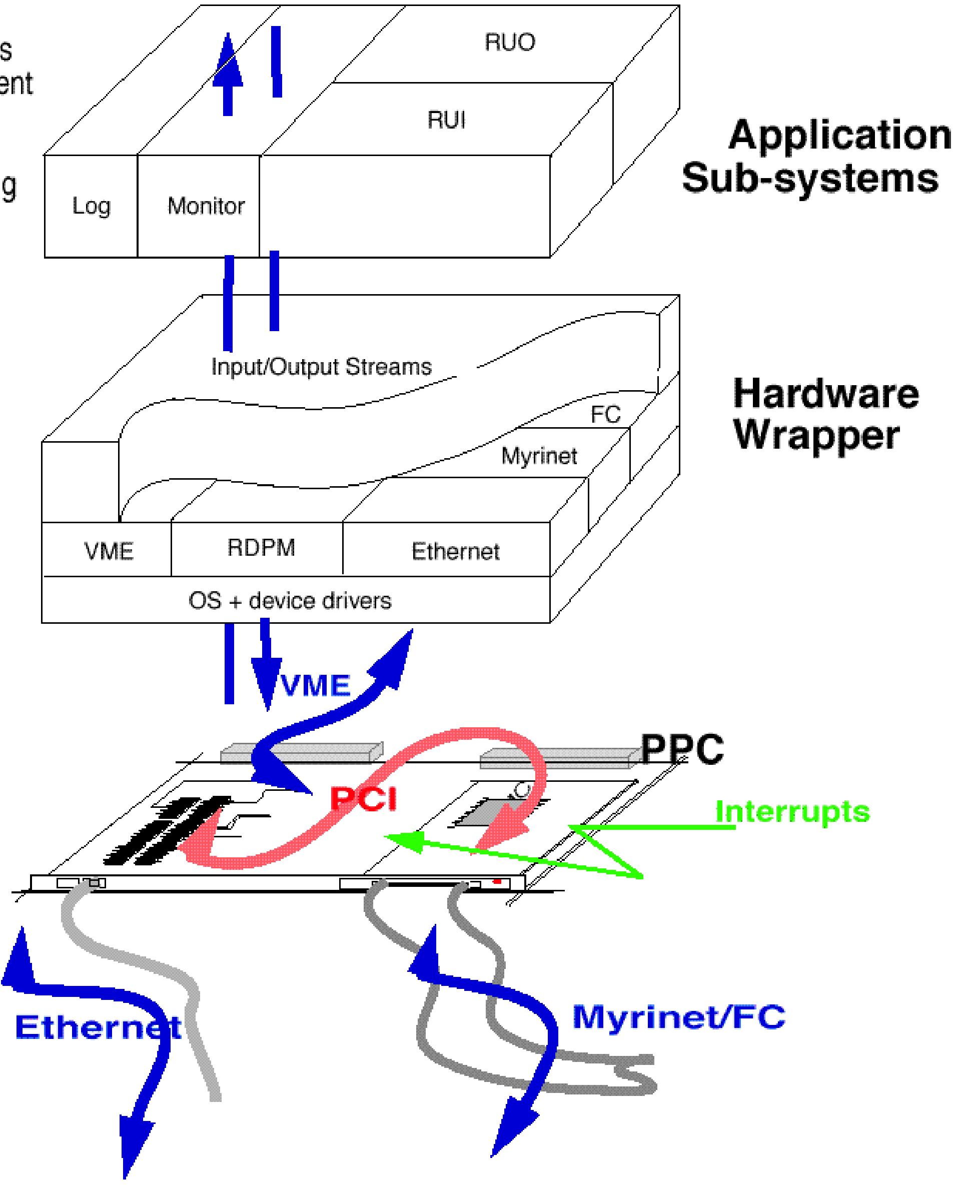
The full RU functionality has been implemented in a software package, written in C++ and running under VxWorks. The software modules are structured so as to run in multiple configurations varying from a single CPU performing the RUI, RUO, RUM and RUS tasks to mixed structures with separate CPUs for RUI and RUO and a FPGA-RDPM as a RUM unit. The present version has been used in the DAQ demonstrator at CERN driving Fiber Channel and Myrinet output interfaces with Ethernet as fast control port. More information on the architectural breakdown of the Readout Unit can be found at http://cmsdoc.cern.ch/ftp/distribution/tridas/4.Talks/appendix.pdf

Readout Unit VxWorks based

System sub-system Layout

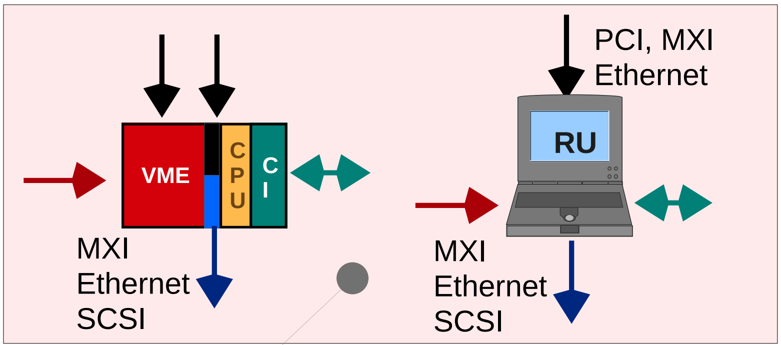
- handle trigger L1/2 requests
- event fragments management
- support traffic shaping
- event fragments Rx, Tx
- configuration and monitoring
- logging
- test and measurements

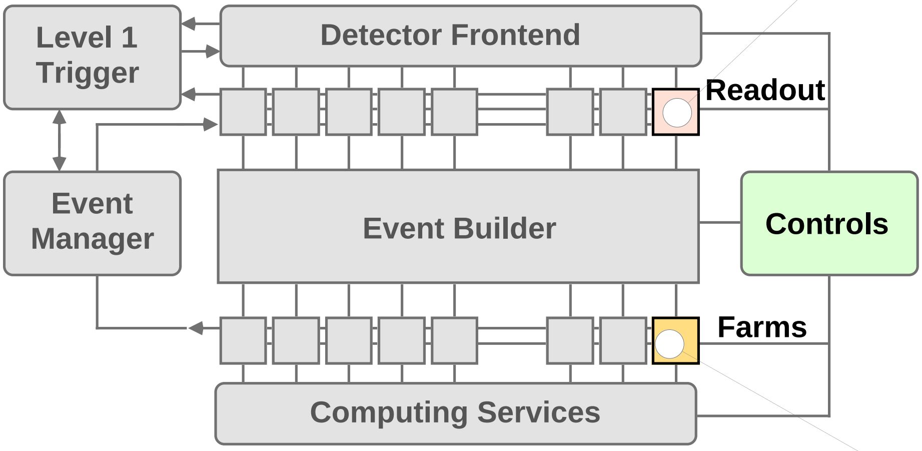
- interrupt handling
- synchronization
- device control
- data Tx, Rx
- memory management
- test and measurements



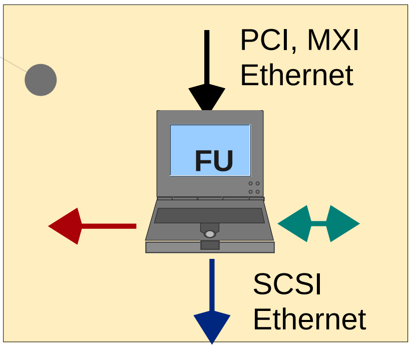
RU development for small DAQ systems (RUVx)

CPU based **READOUT UNIT** (VxWorks OS). Two Configurations:
1) VME-CPU, IN=VME, OUT=Ethernet/MXI
2) PC-CPU, IN=PCI, Out Ethernet/PCI





Control network based on Ethernet. **RUN PANEL** via Web pages



PC/WS based **FILTER UNIT** (UNIX OS).
Event fargment IN=Ethernet/MXI/PCI, OUT=SCSI/RDS

RU based DAQ Systems

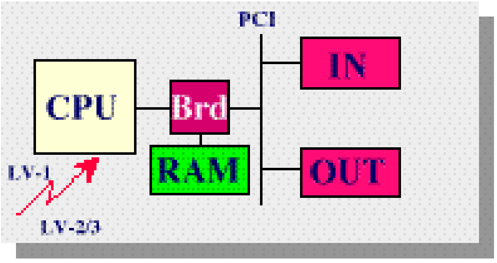
The RU can be seen as a standalone module performing all the basic functions of a classic data acquisition. It can then find applications different from the final CMS DAQ. In particular it can be used as building block of daqs aimed to smaller scale experiments or detector beam tests.

According to the application needs different form factor can be used to implement the RU
- PMC
- Desktop PC
- CompactPCI

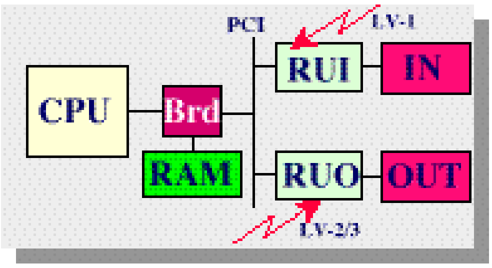
The application needs drive also the choice of the RU components

PERFORMANCE

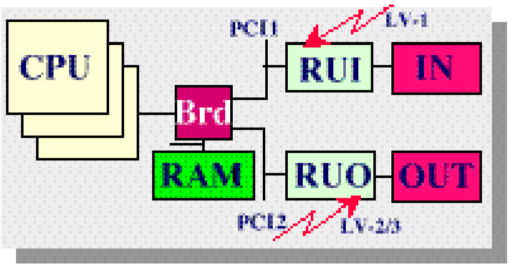
Sofware based RU (vxRU); e.g. VME/PMC, Desktop 3U/6U CPCI



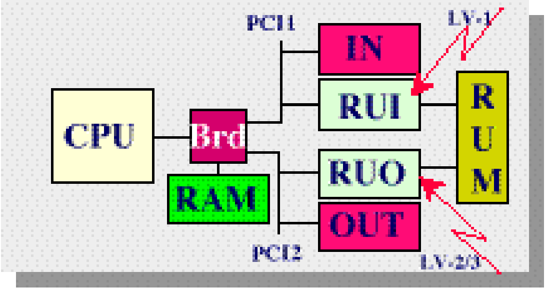
Software based RUM with RUI/RUO IOPs. e.g. VME/PMC, Desktop, 3U/6U CPCI.



as above, but with double PCI and multi CPU booster. e.g. Desktop



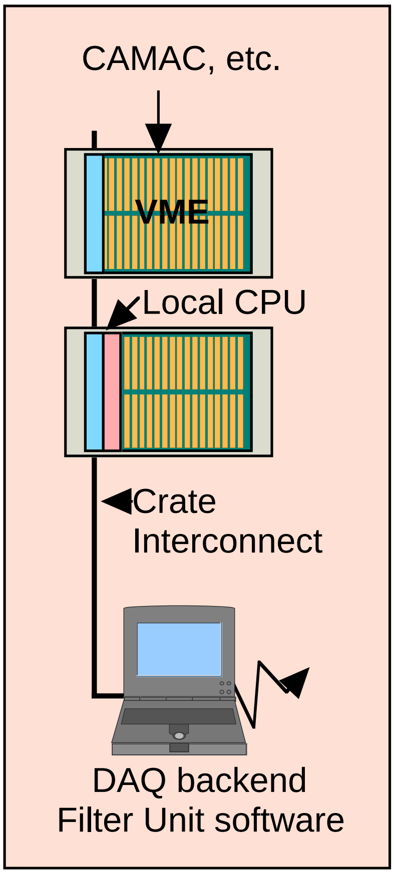
FPGA based RUM, RUI/RUO IOPs, 2 PCIs; host CPU as RUS. e.g. 6U CPCI, CPCI+Desktop



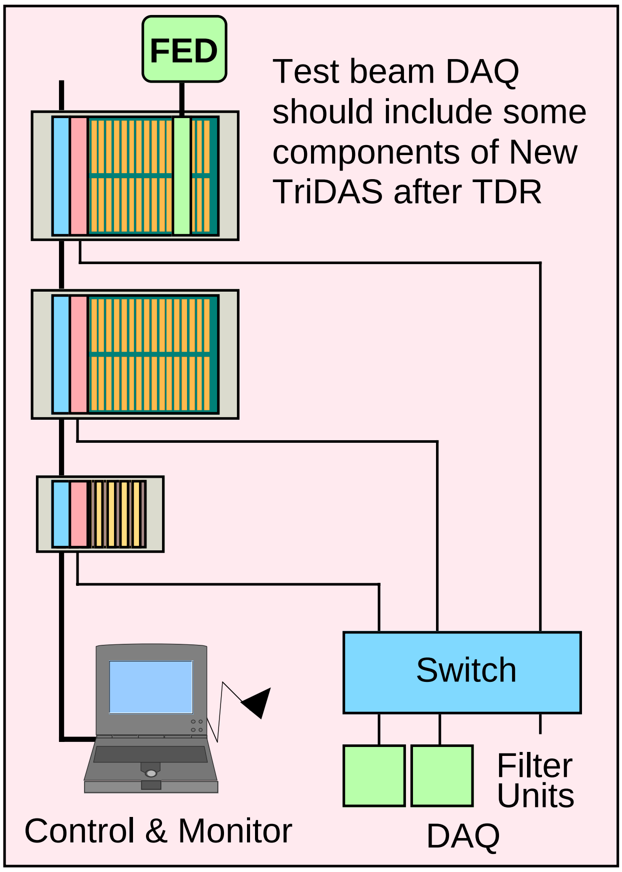
G. Maron, TriDAS meeting, CERN 9-10 November 1998

Test beams

Test beam short term

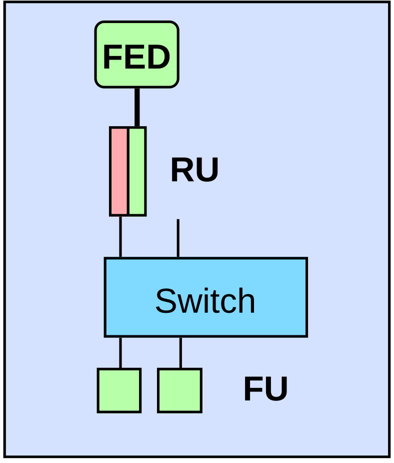


Test beam after TriDAS TDR



During the transition phase, test beam DAQ systems should make use of the most generic components such as Desktops, PCI-VME crate interconnection, standard CPUs, popular OS and LabVIEW for test and controls

TriDAS prototypes



The TriDAS current work mainly addresses the DAQ final design issues by :

- prototyping and testing hardware and software functions
- evaluating data communication technologies
- simulating system architectures

A medium size DAQ demonstrator should be implemented by the end of 1999

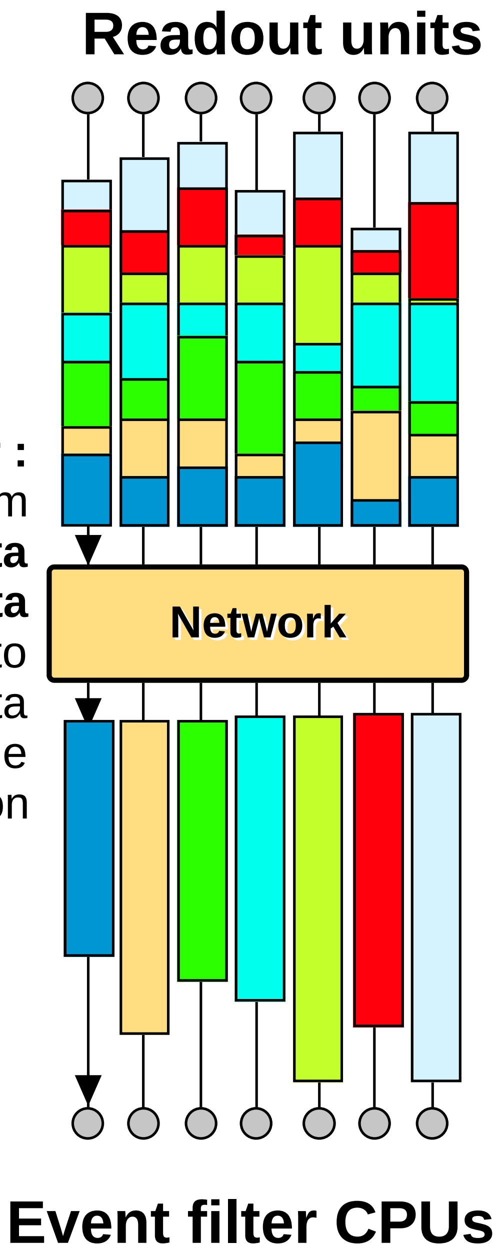
DAQ demonstrators 98 milestones

Event Builder test-bench: A 4x4 Event Builder based on multiple switch technologies and inputs and outputs (RUs and FUs) resembling the current module developments as closely as possible.
- **Switch:** both Fibrechannel (ANCOR) and Myrinet switches will be utilized
- **EVB prototype simulation**: it will provide a functional model of the 4x4 EVB setup.
- **RCN/FCN networks:** implemented via external Ethernet switches. Both TCP/IP and light protocols will be used.
- **Filter Unit:** Farm monitoring and control

Report on 98 December milestones

On the FibreChannel (FC) side, a full set of system measurements is now available. First conclusions are that the control networks (RCN and FCN), implemented through Fast Ethernet, are the main sources of the current performance limitations. On the switching front, the Tachyon chip (found on the FC Network Interface Cards) presents limitations in achieving the full theoretical maximum FC bandwidth due to its inability to handle (simultaneously) fragments from different sources. While this is not a serious limitation for the typical inter-computer communication needs, in the Event-Builder environment, which relies heavily on the simultaneous traffic between the roughly 500 sources and 500 destinations, this feature is highly undesirable. The CERN demonstrator has also seen the arrival of the Myrinet switch and corresponding PCI/PMC interfaces. The first impressions from this new technology are its quick deployment that has resulted in first measurements of the point to point latencies (and throughputs).
See a summary of DAQ demonstrator results at
http://cmsdoc.cern.ch/ftp/distribution/tridas/Meetings/TriDAS.weeks/98.11.09/DEM-DS.pdf .

Event Builder



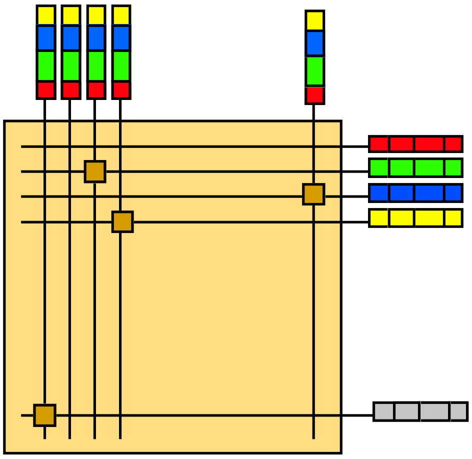
Event builder : Physical system interconnecting **data sources** with **data destinations.** It has to move each event data fragments into a same destination

Event fragments : Event data fragments are stored in separated physical memory systems

Full events : Full event data are stored into one physical memory system associated to a processing unit

Event building by switches

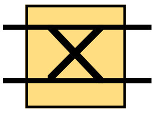
1) Cross-BAR



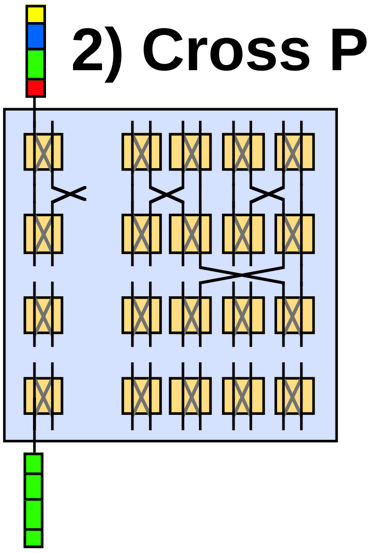
Limited in size (up to 64 ports). Building block of many products. Protocols:
- External control (e.g. barrel shifter)
- Circuit switching (Node autorouting)

The maximum switch load for random traffic is about 62%

2) Cross Point FABRIC



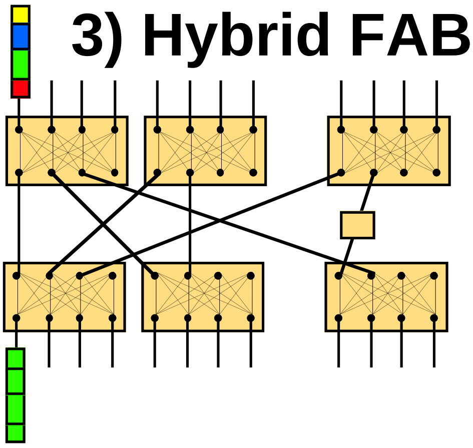
2 x 2 switch unit
basic component



Large size (> 1000 ports) attainable
High level protocols:
- Packet switching
- Traffic shaping (rate division, backpressure)
- Multipath structures, Internal buffering

The maximum switch load with traffic shaping can be of the order of 80%

3) Hybrid FABRIC



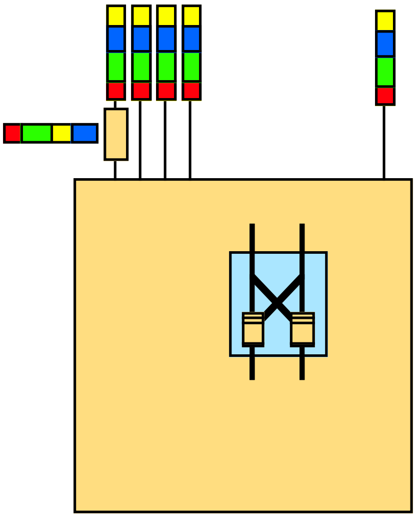
Large size (> 1000 ports) attainable
extension of crosspoint fabrics

The maximum switch load depends on the traffic shaping algorithms and the global system internal resources (multi-path, buffer size etc.)

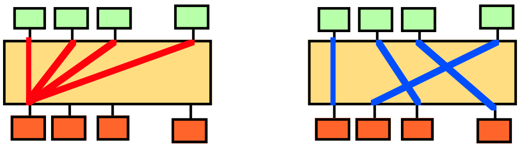
Switch load (throughput) and protocols

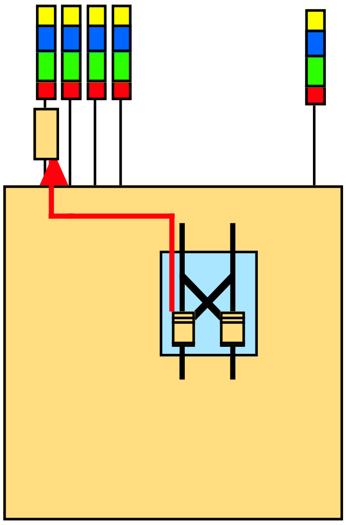
Switch load = (Total effective bandwidth)/(Bisection bandwidth)
Switch Bisection bandwidth = the sum of the bandwidths of all the channels crossing a line separating all input ports from output ports

The attempt to move data to the same port or internal node creates a CONGESTION with possible loss of data. Congestion can be reduced (switch load increased) by special protocols (at the cost of more I/O logics and buffering)

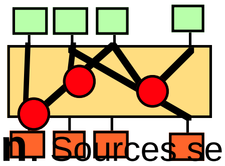


Traffic shaping (e.g. Alcatel)
- randomized multi-event data input to the switch.
- design the switch node buffers such to have an acceptable **cell loss** probability
- not simple adapter level
- good aggregate bandwidth (load up to 80%)





Back pressure (e.g. AT&T Phoenix, IBM PRIZMA)
- Inhibit previous stage data sending when buffer are full
- no cell loss
- simpler adapters
- low aggregate bandwidth (load 30%)



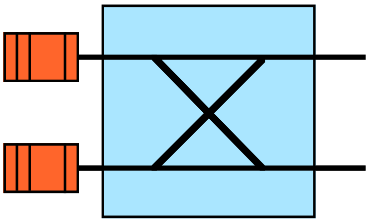
Rate division. Sources send cell in a round-robin manner
Barrel shifter. External signal synchronizes sources

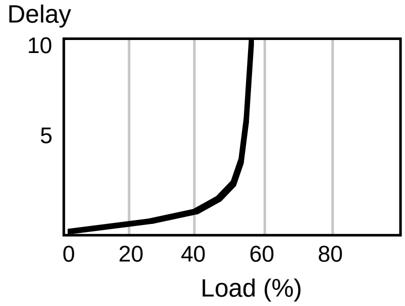
ATM switch fabric main parameters

- Latency: the delay of data through the switch
- Switch throughput: the usable fraction of total bandwidth
- Switch scalable, modular and upgradable
- Cell loss probability and fault tolerant

Buffering methods

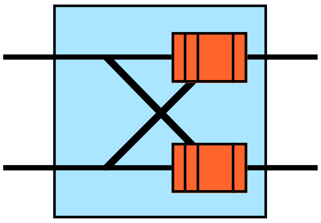
Input Queueing

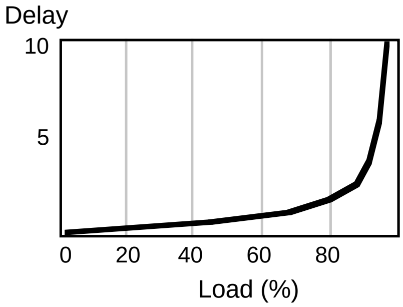




Cells are hold at the input port until output is free. A waiting cell may "block" others

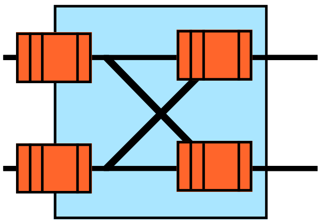
Output Queueing

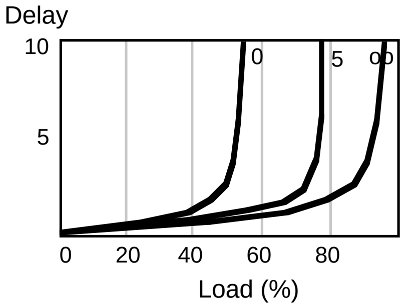




The best performance (delay, throughput) way but more complex circuit

Combined Input Output Queueing





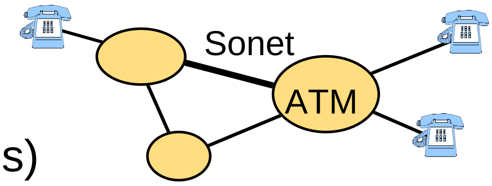
Most efficient use of memory, scalable, but complicated architecture

Switch technologies

NETWORKS and TELECOMMUNICATIONS

- Asynchronous Transfer Mode (ATM)

- Link bandwidth: **155, 622**, 1244, 2488 Mb/s
- Packet switch based on small cells (53 bytes)

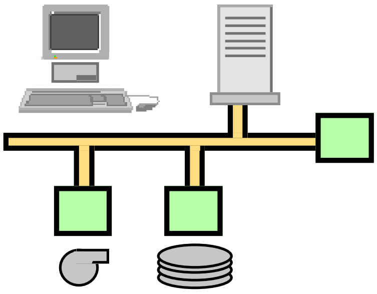


≈ 10^6 nodes

PERIPHERAL NETWORK. (Massive Parallel Storage Applications)

- Fiber Channel System (FCS)

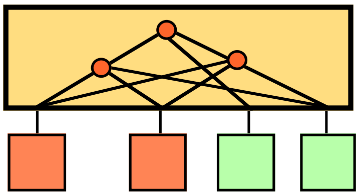
- Link bandwidth: **133, 266, 531, 1062** Mb/s
- Frame based data delivery (2148 bytes)



≈ 10^3 nodes

HIGH PERFORMANCE COMPUTING AND COMMUNICATION (HPCC)

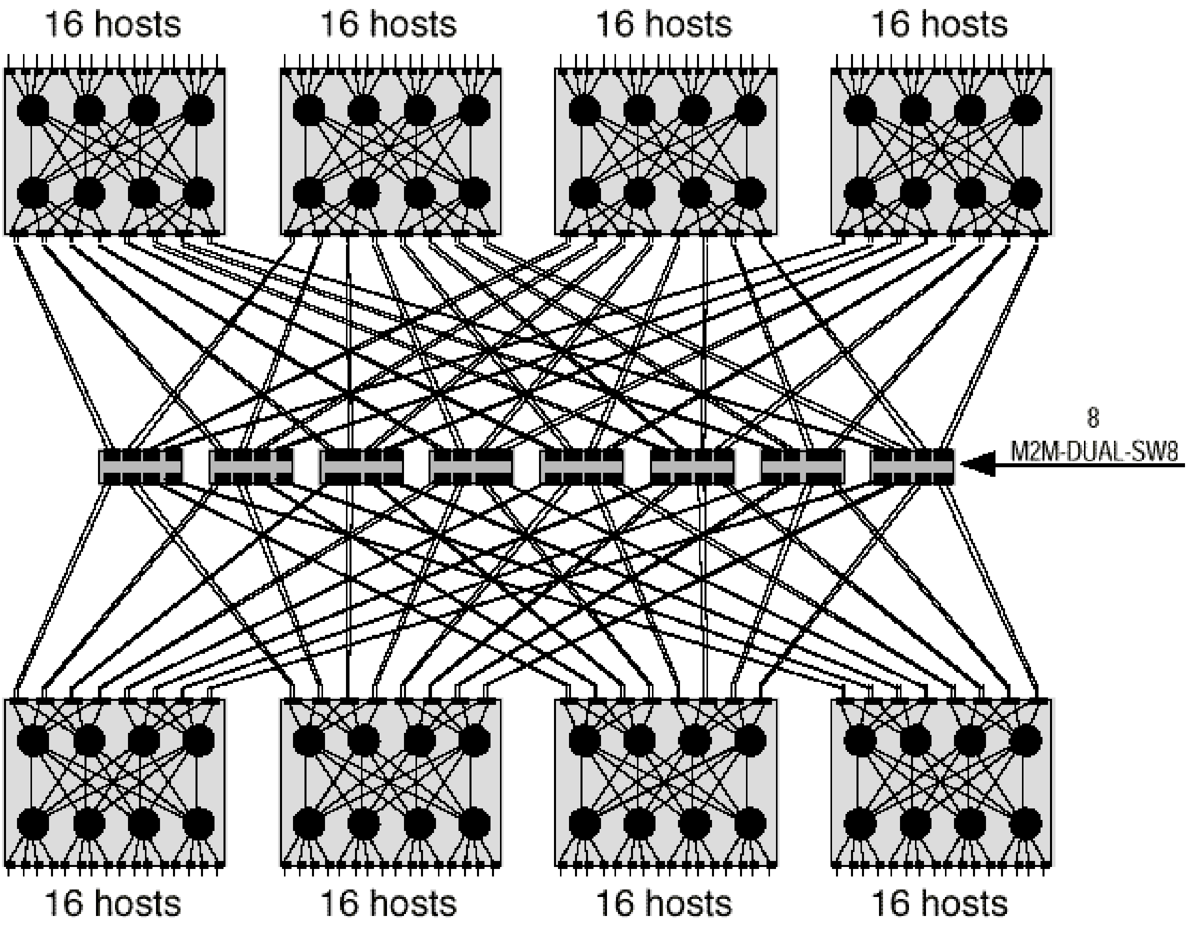
- Mercury RaceWay (160 Mbyte/s)
- **Myrinet (1 Gb/s)**
- **GEthernet**
- Computer manufacturers proprietary switch (Application Strategic Computing Initiative ASCI)



≈ 10^2 - 10^3 nodes

128 ports Myrinet network

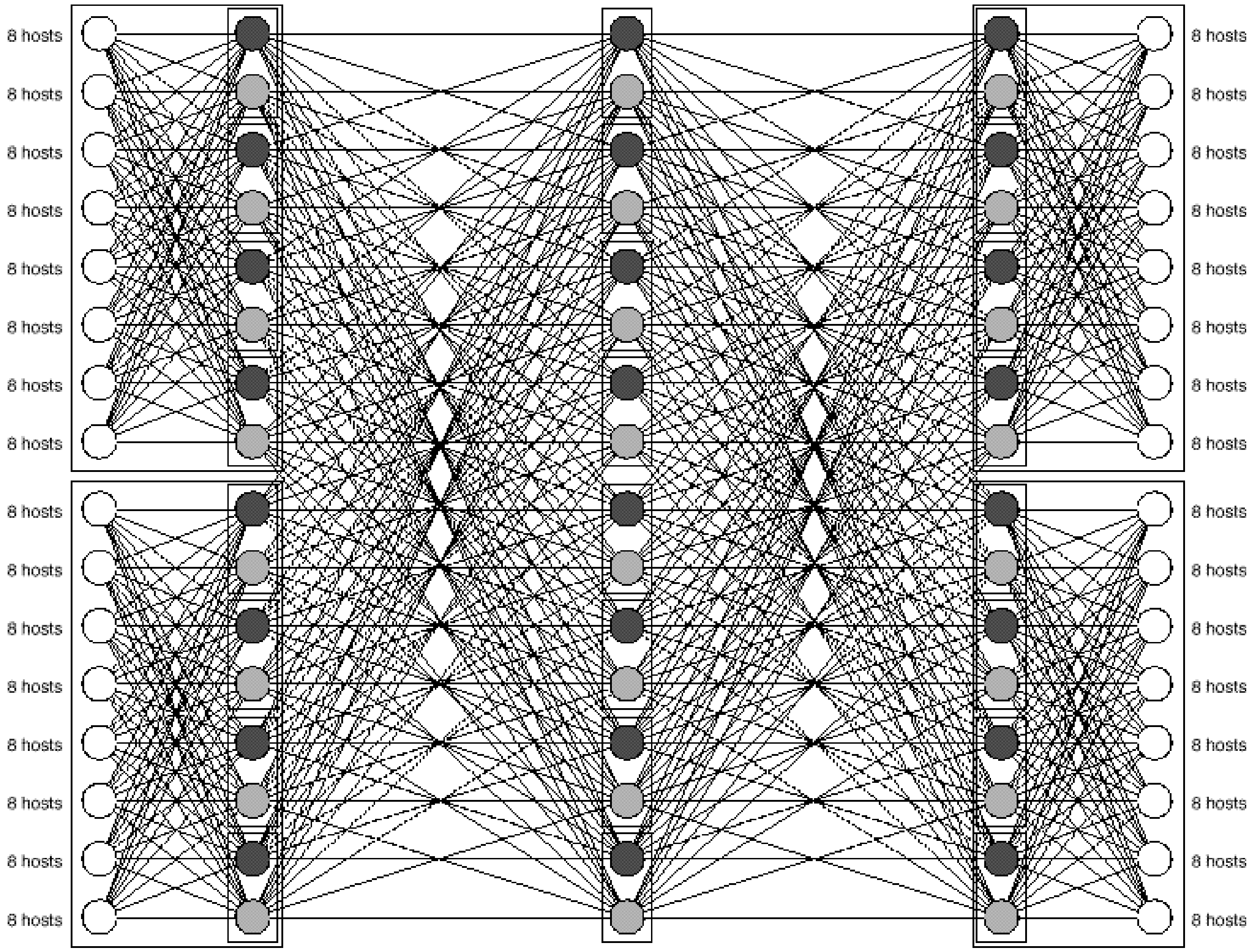
N = 128, BBW = 64 (full), D = 5, D_{av} = 4.7



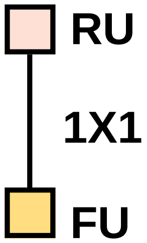
This multi-stage graph with 5 stages provides the full BBW of 64 links (81.92 Gbits/sec).

Qty	Component	Each	Total	Subtotals
128	Myrinet-SAN/PCI interface	$1,300	$166,400	
128	5-foot SAN cable	$140	$17,920	$184,320
8	Octal 8-port Myrinet-SAN switch	$6,000	$48,000	
8	Dual 8-port Myrinet-SAN switch	$2,000	$16,000	
64	5-foot SAN cable (switch/switch)	$140	$8,960	$72,960
(Total cost per host = $2,010)			Total:	$257,280

256 ports Myrinet

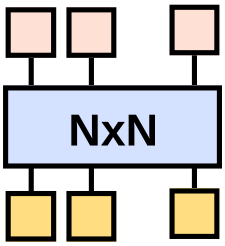


DAQ demonstrators



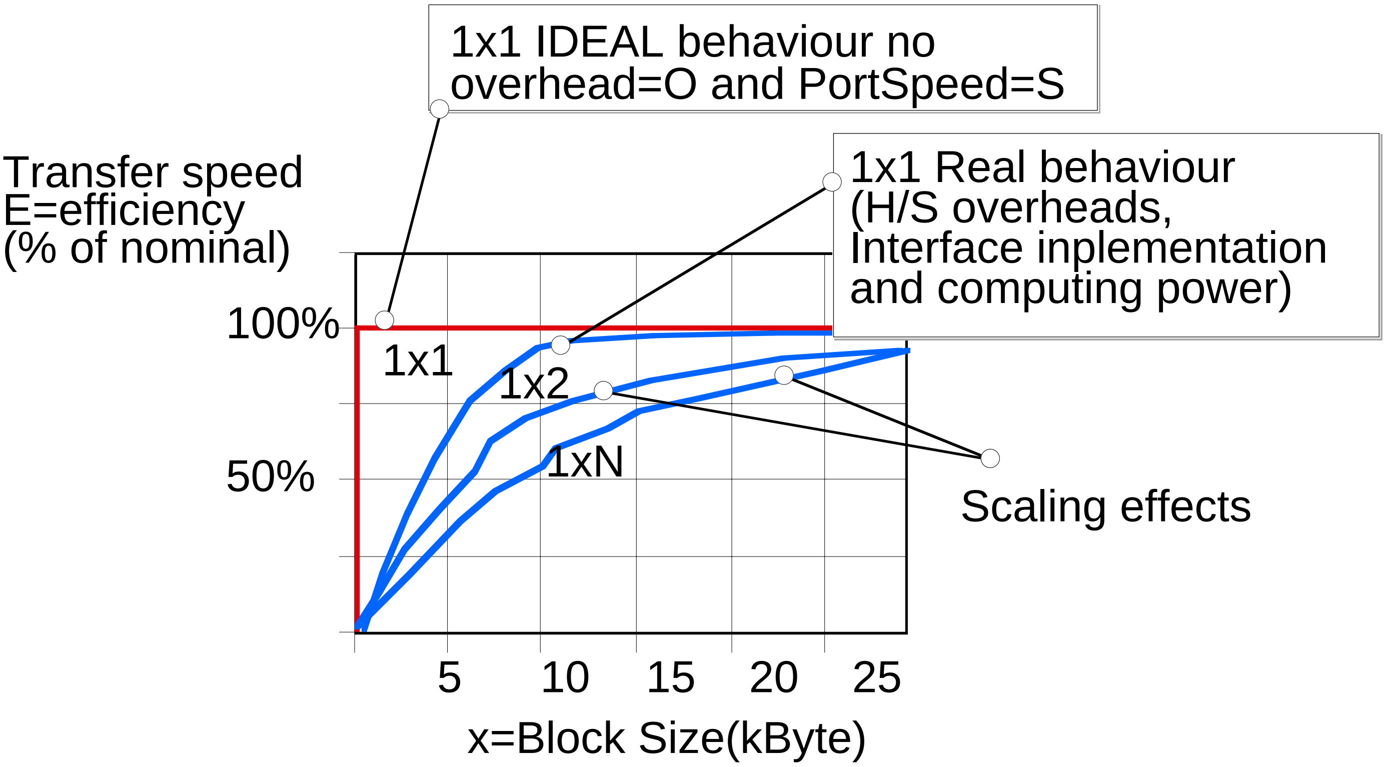
1) DATA LINKs :

Measure point to point link speed
(ATM, FCS, Myrinet, Ethernet)



2) EVENT BUILDER/SWITCH:

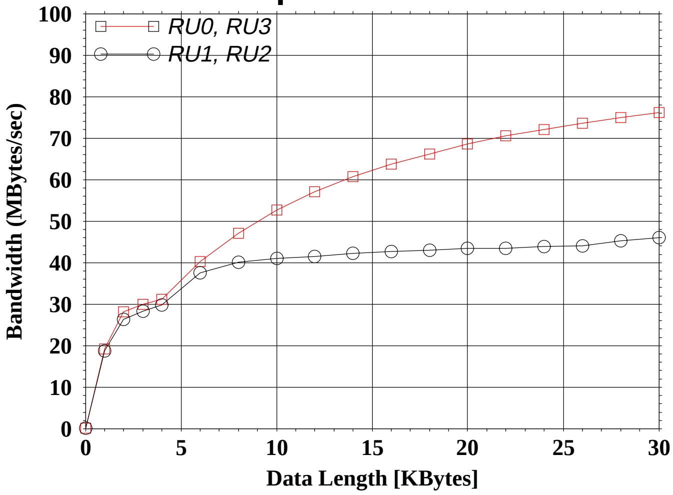
Measure scaling effect with multiple RUs and different protocols. Study deviation of:
PortSpeed$_{NxN}$?=? N x PortSpeed$_{1X1}$

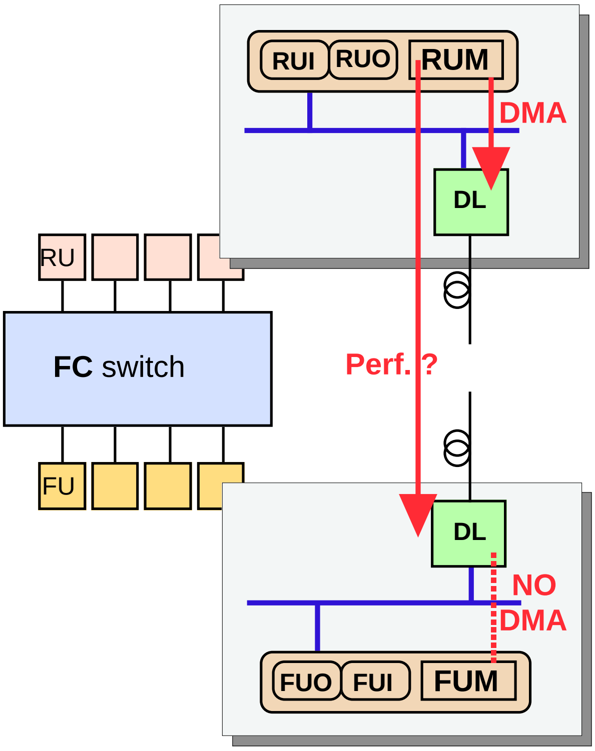


$$E= x/(x+O \bullet S)$$

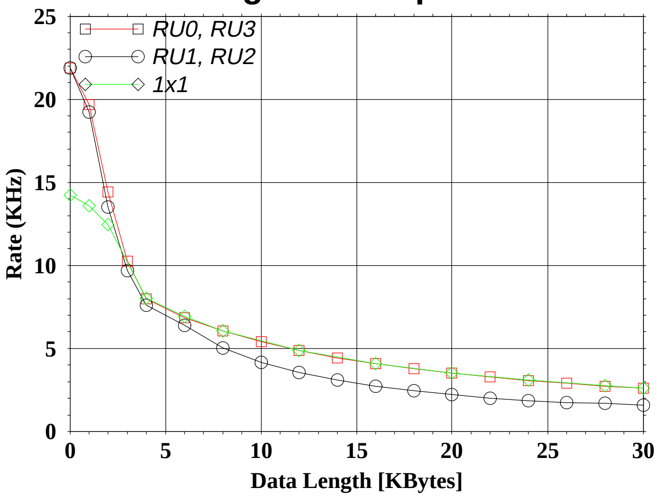
FC Point to Point tests

RU Output Bandwidth

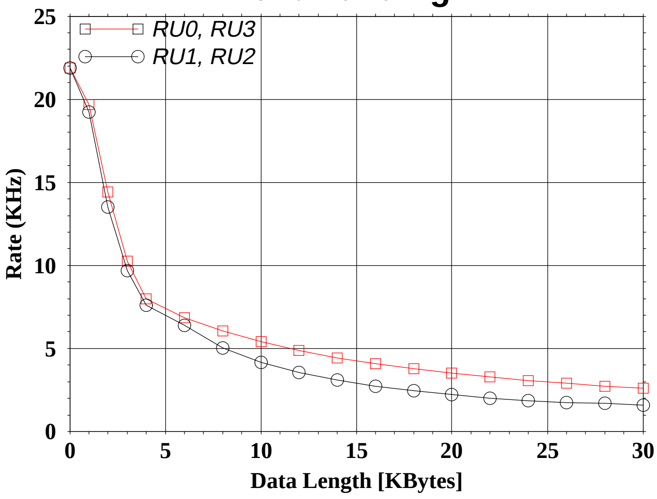




RU Fragment Output RATE



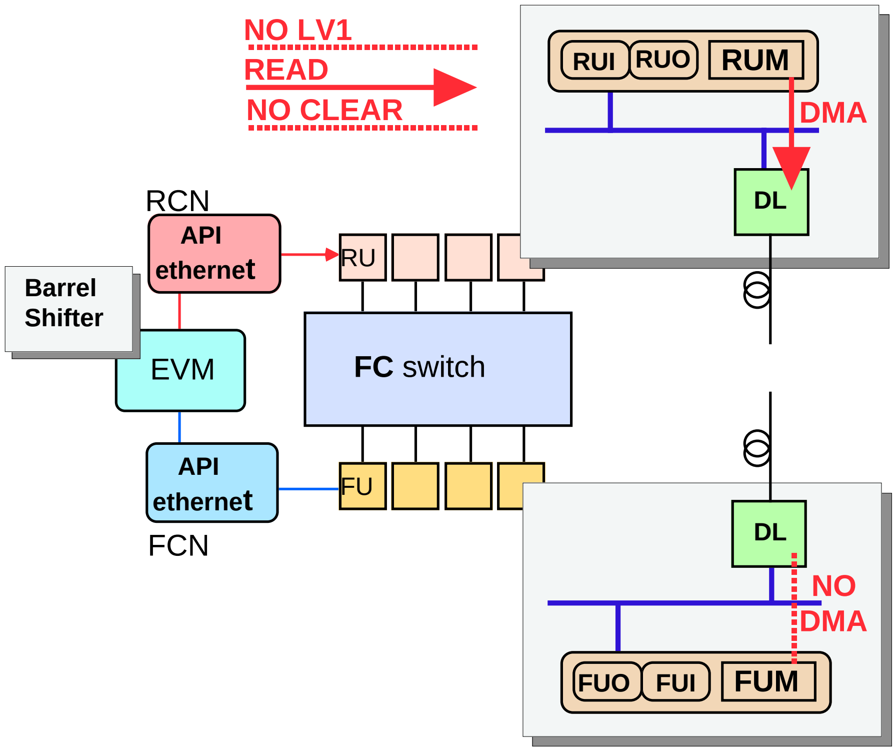
RU Fragment Output RATE compared with 1x1 Event Building RATE

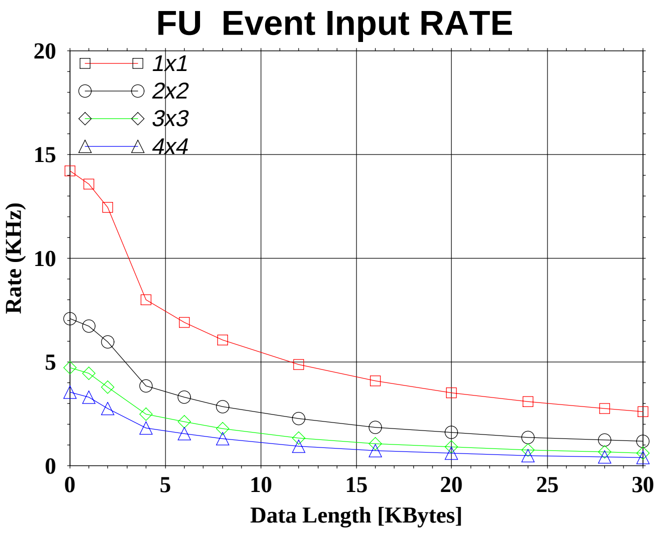


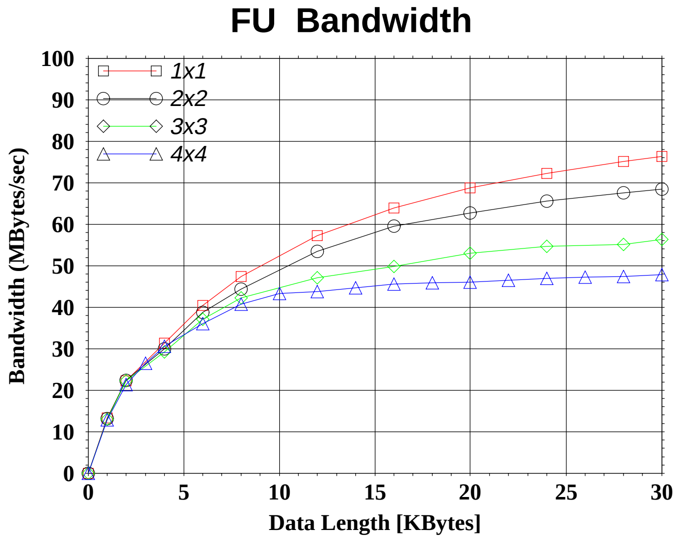
- RATE (Bandwidth) depends on position of the RU on the switch

- RU saturation (>3K)

Scaling with FC switch



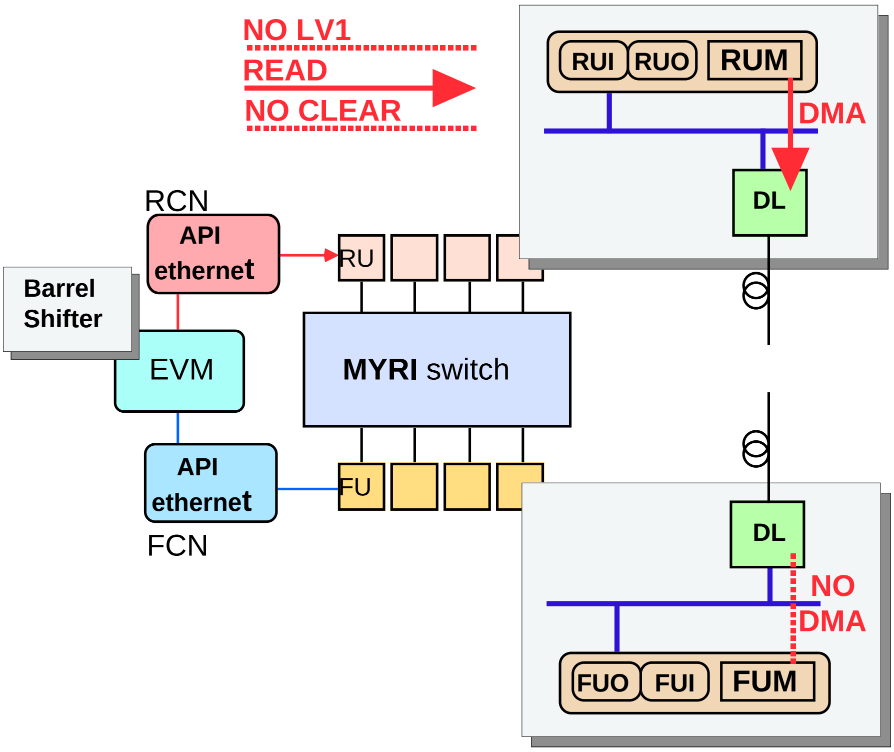


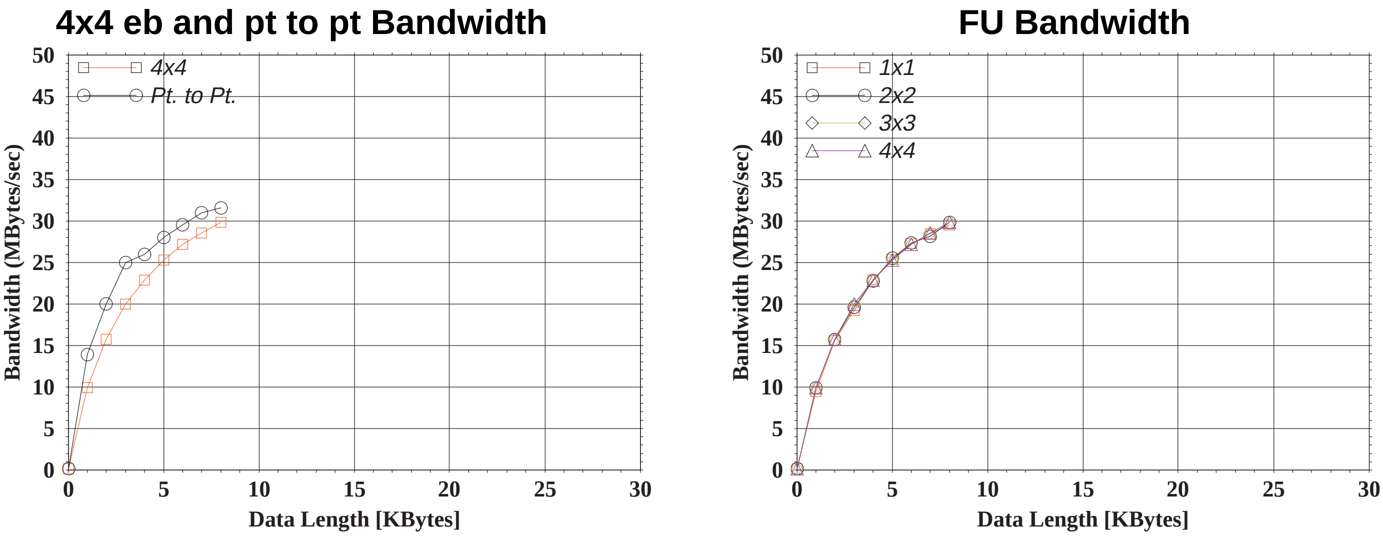


NO scaling:

- switch is not "uniform"
(bandwidth depends on the position of the link on the switch)

EB with Myrinet switch





4x4 Bandwidth

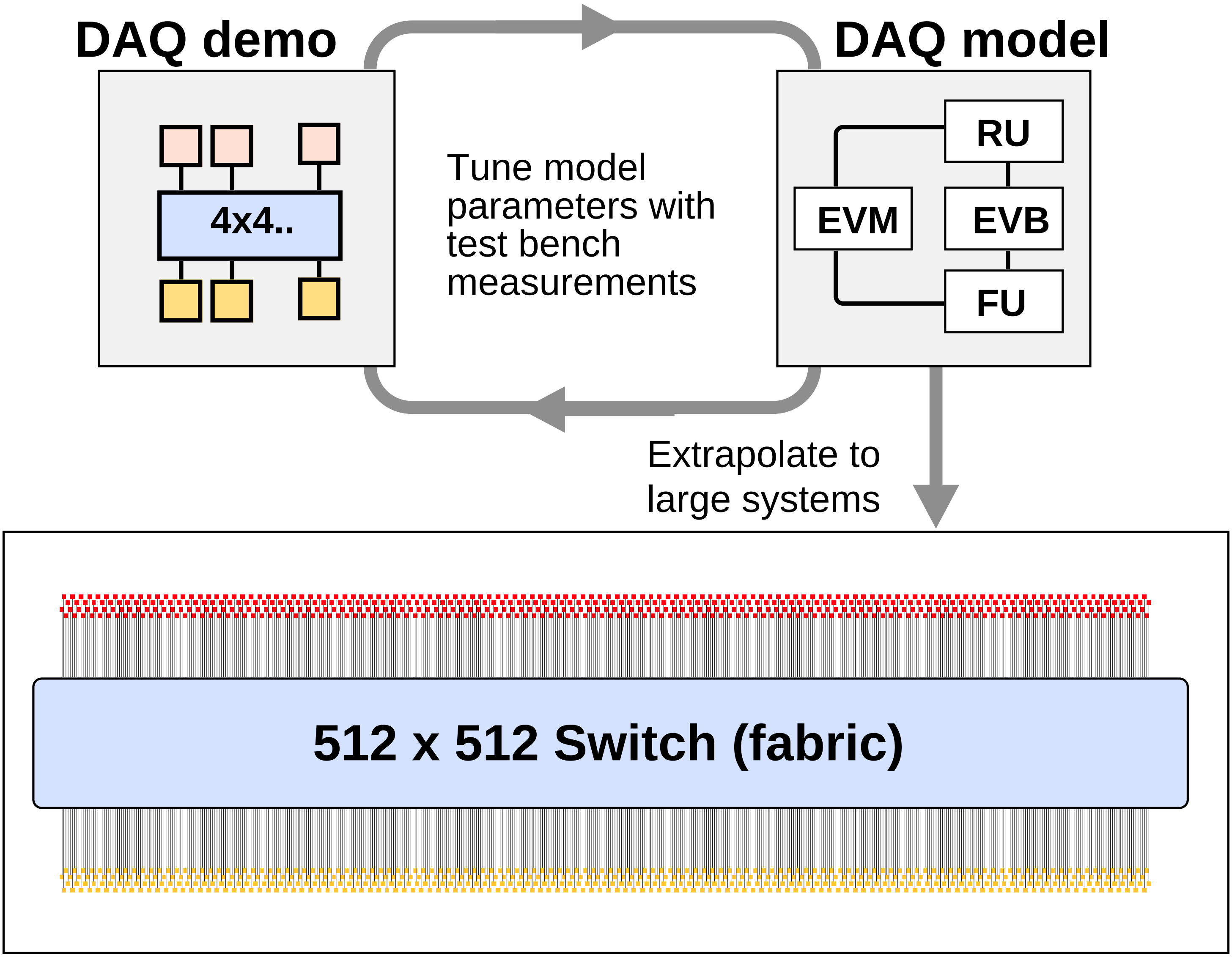
- 8K software limitation

Event Building with Myrinet switch

SCALING (0 up to 8 KBytes)

DAQ simulation

• Building the CMS DAQ System requires an extensive study of different technologies and system architectures
• Small scale prototypes are build in order to evaluate technologies
•System simulations are expected to reproduce the performance numbers obtained by the prototypes
• Once the simulations are validated and verified, we can learn how the final system will behave without having to build it first
•System simulations of that scale are as complex as building a prototype system. However, simulations give much more flexibility



Tune model parameters with test bench measurements

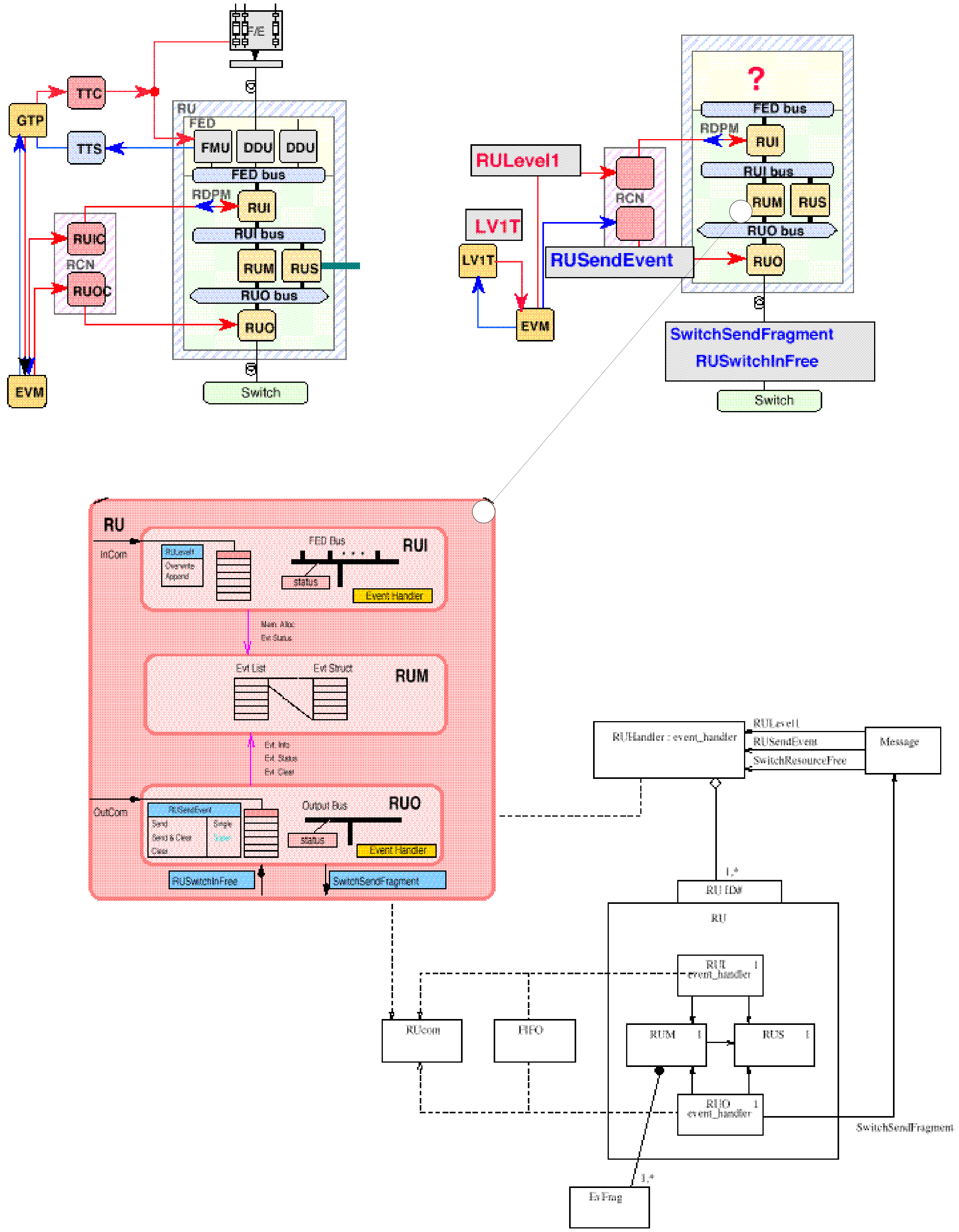
Extrapolate to large systems

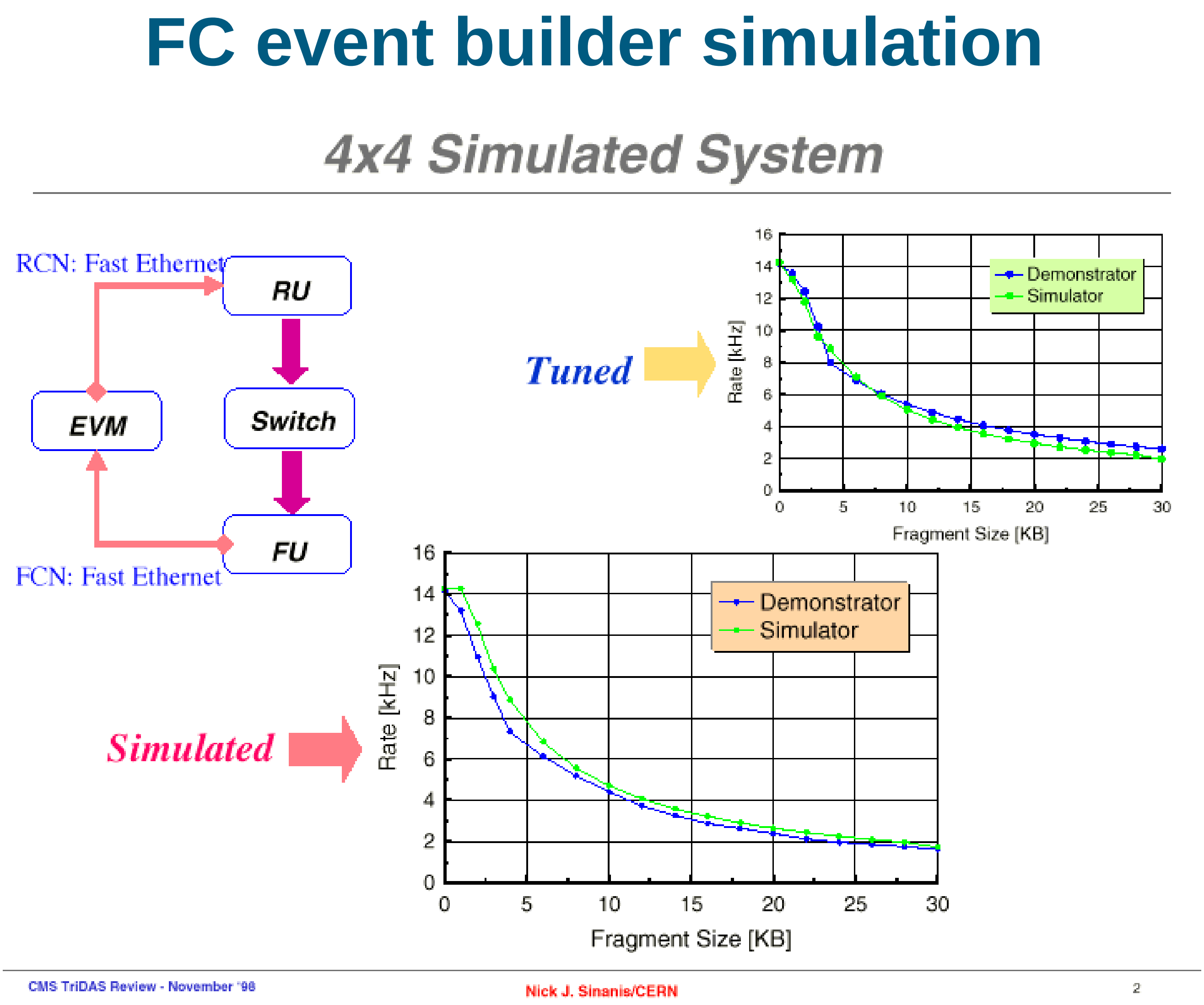
- FrontEnd data flow simulation (FE, Links, Buffers) [models in Foeresigth and VHDL]
- Event builder simulation (RU,FU,Switch, EVM,protocols) [Model in C++ using Computing Network Class Library (CNCL)]

Example: RU model

Actual RU Device

Modeled RU





FC event builder simulation

4x4 Simulated System

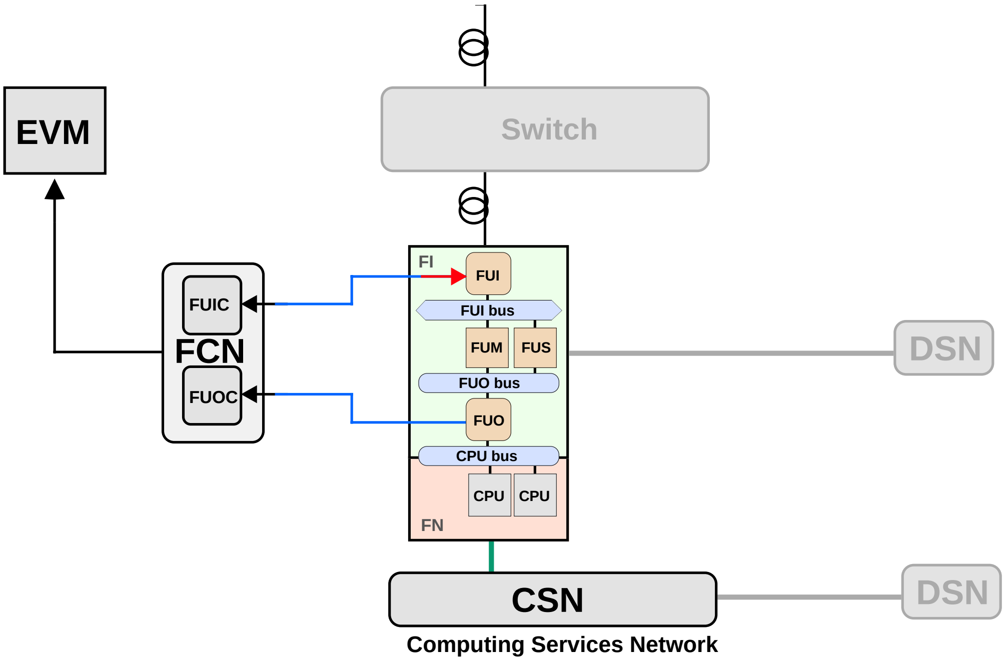
CMS TriDAS Review - November '98

Nick J. Sinanis/CERN

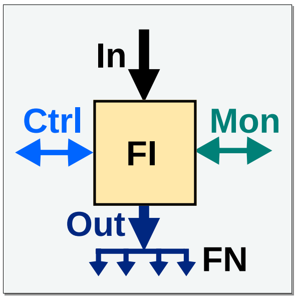
2

- A modular simulation system of the CMS DAQ has been developed
- We have designed a suitable tool to study DAQ system scaling issues
- Verification with the DAQ demonstrator, still goes on.
  The present results show a good match with basic speed measurements
- Invest more on Simulator Validation using Message Sequence Charts generated by the simulator engine

Filter Unit



Farm interface (FI, SFI)
Farm controls
Farm Node Applications framework
- Software framework
- High Level Trigger algorithms
- Mass storage and OODB



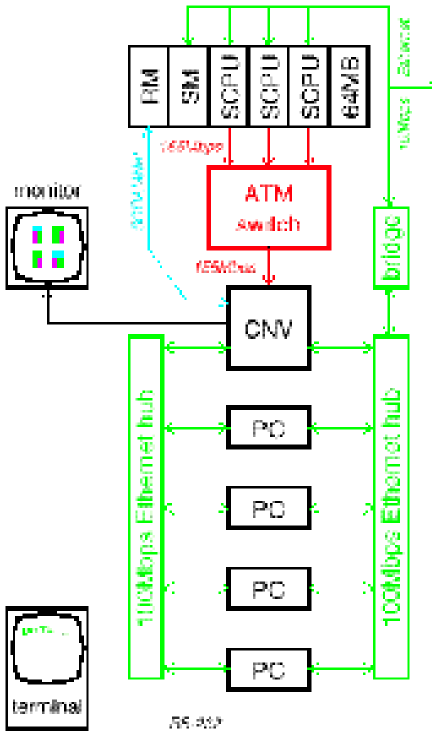
IN (link)	**200 MB/s**
OUT (bus/link)	**200 MB/s**
Ctrl (network)	**≥10 MB/s**
Mon (network)	**≥10 MB/s**
MEM	**≥256 MB**
CPU	**~ 10 TeraOPS**

Event Builder/Filter testbench for ATM

CDF Converter Nodes

CDF converter nodes serve the same function as our SFI.

Input fragments on 155 Mb ATM.
Output events on two 100 Mb Ethernets.
Distribute to group of CPUs using hubs.



PC farm cost effective.
200 Mhz P6 converter, Linux.

Dual Fast Ethernet output for standard, cost-effective links. (commodity)
Almost sinks full ATM rate.

Affordable, commodity, upgradeable...

Tridas Review Nov. 9, 1998 James G. Branson

Server's architecture (e.g. SUN UltraSpark, IBM, Compaq, Intel ...)

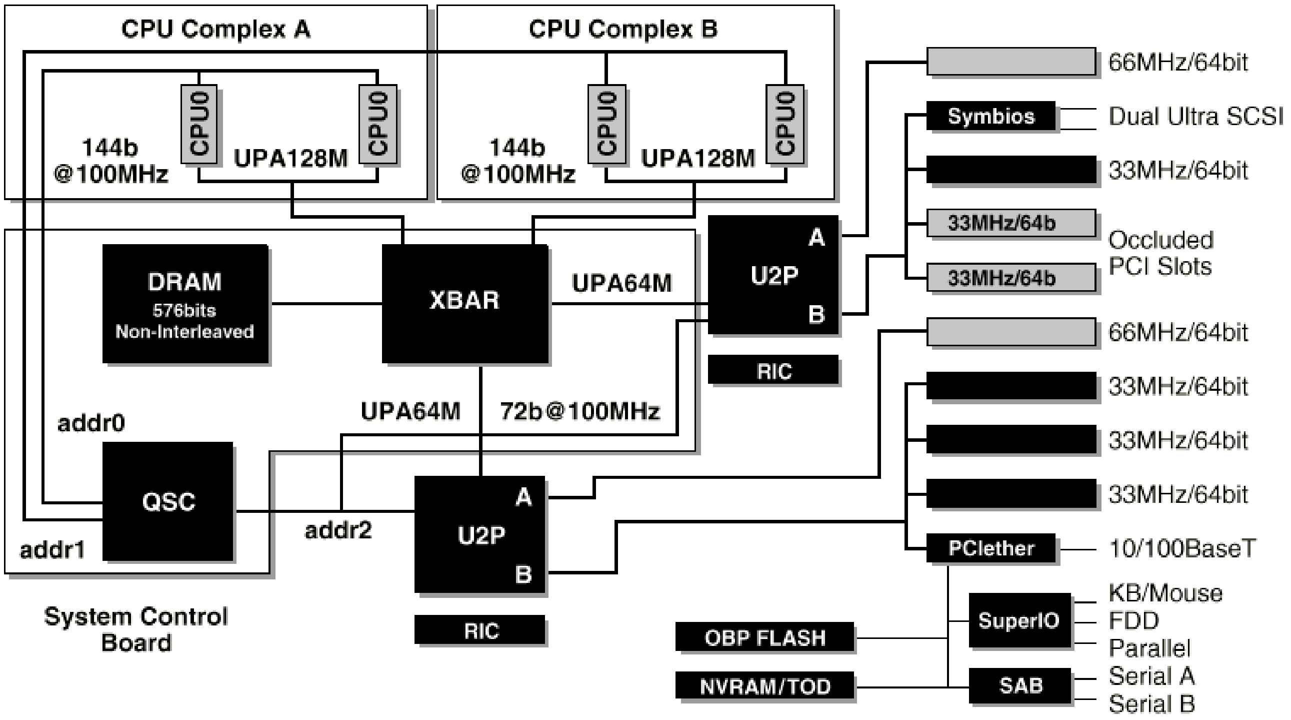
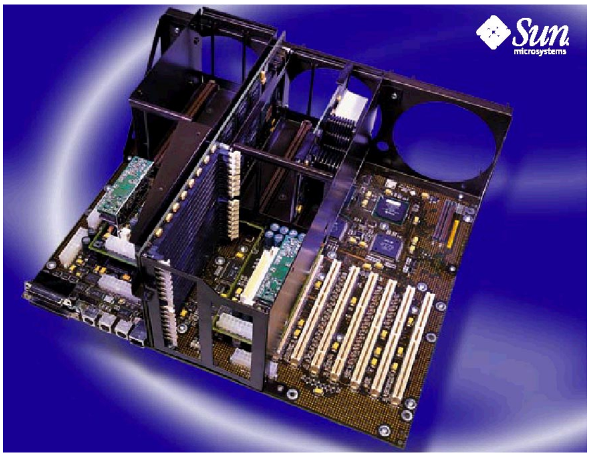
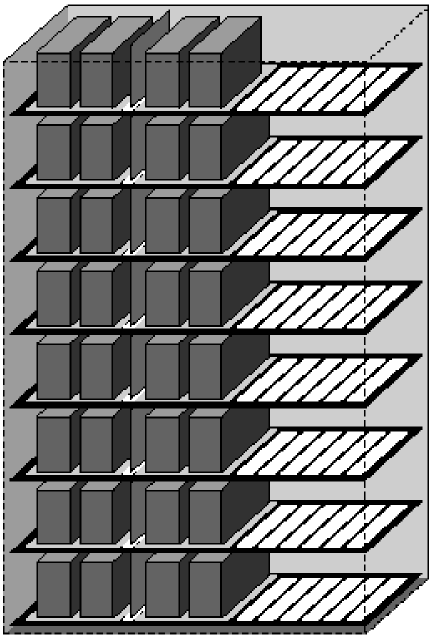


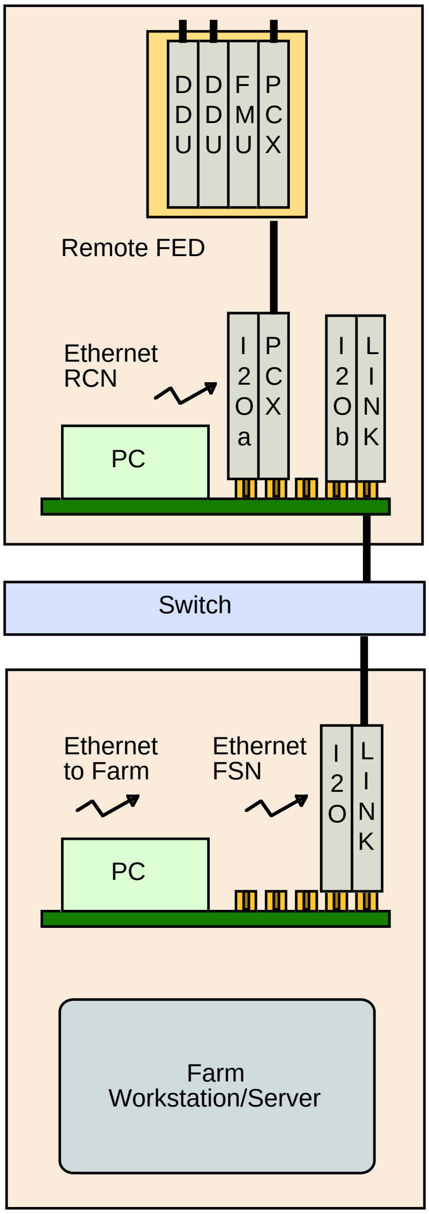
Figure 4. SPARCengine Ultra AXmp Schematic View





up to 40 UltraSPARC processors and 60 PCI cards per 42U rack.

Desktop DAQ column



Desktop RU. Description

RU system based on desktop PC. The CPU runs the RDPM event handling logics, the data buffering and the monitoring (RUM and RUS). The FED system is readout via a PCI-to-PCI connection (PCX) driven by an I2O board (RUI), a second I2O (RUO) drives the output link reading data from the PC memory (RUM) via PCI. EVM commands (RUIC/RUOC) are received via ethernet port in I2Oa(RUI) and I2Ob(RUO). The PC ethernet interface is used as DSN port.

Desktop FU. Description

FU system based on desktop PC. The event data fragment input link is driven by an I2O module. The CPU runs the event data memory handling logics, the data buffering, the unit monitoring and the farm communication logics (FUM, FUS, FUO). The connection with the processor farm is implemented by ethernet (or a PCI-to-PCI extension).

High Level Triggers 98 milestones

HLT results on single-e,μ triggers

- First results from simulation and HLT studies; evaluation of rates relative to Level-1 trigger with and without tracking information.

Report on 98 December milestones

The production of large background samples of events passing the calorimeter and muon-based Level-1 triggers has been started. Single-jet background rates folded with particle-level QCD jet rates are used to obtain the input "Level-2" rates and to measure the rejection of various additional requirements by the HLT (e.g. the presence of a charged track in a Level-1 inclusive electron trigger). A preliminary result is the confirmation that about half of the high muon rates at Level-1 are indeed due to heavy-flavor decays, indicating that sophisticated topological and physics-based algorithms will be needed to further reduce the rate after the Level-1 trigger.

The study on track reconstruction algorithms and the role of different tracker planes have been initiated as well. Preliminary results indicate that for electrons one should start from inner layers (e.g. pixels) while for muons, starting from outer detector layers (MSGCs) should be adequate.

High Level Triggers: Current work (summary)

Currently working on

- **Inclusive electron trigger**
- **Inclusive muon trigger**

  for both: major task is to generate the backgrounds

Requirements

- **We need to get full simulated background events, not just an estimate of the rates (so one can apply filtering algorithms beyond Level-1)**
- **For the time being: use CMSIM (115)**

  few exceptions (e.g. ECAL C++ reconstruction)
- **Clearly, tracker information will be used at some point. Current plan: get Level-2 rates using (a) detector-only quantities (e.g. calorimeter) and (b) using full offline track reconstruction. True Level-2 rates should be in between** (defines boundaries)

HLT: electrons (I)

Get: probability per jet to pass Lvl-1 e trigger

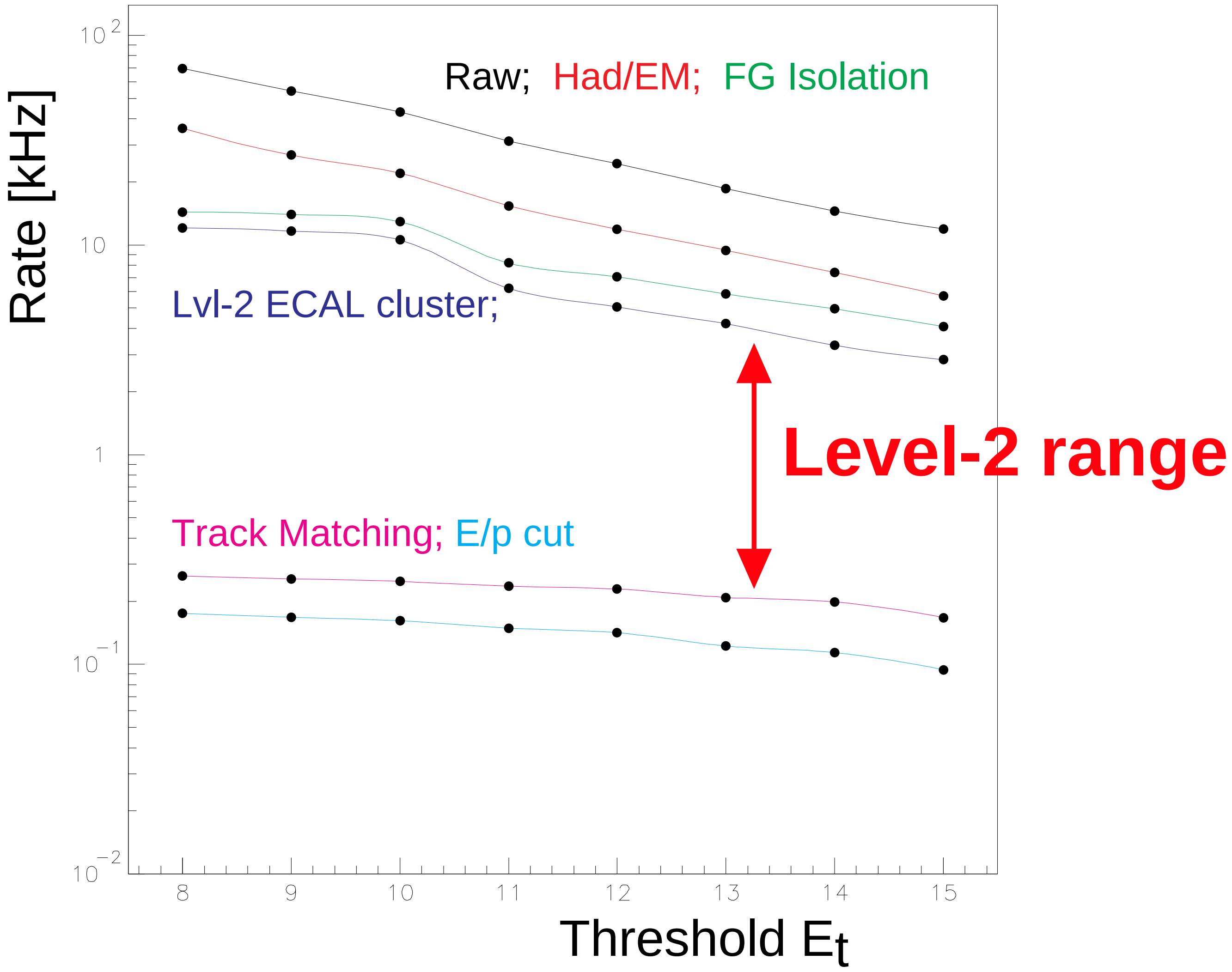
- **Push single jets through simulation (including Lvl-1)**
- **Measure efficiency for passing Lvl-1 vs E_t(part-jet)**
- **Convolute $\varepsilon(E_t)$ with jet E_t cross section (and multiplicity) → get Lvl-1 accept rate**

  Example: PYTHIA event with two jets with E_t^1 and E_t^2

  Define: $\varepsilon_i = \varepsilon(E_t^i)$; Probability to trigger = $\varepsilon_1+\varepsilon_2-\varepsilon_1\varepsilon_2$ etc...
- **Cross-check rates with previous Lvl-1 studies (note: only BARREL so far)**
- **Then: apply cluster-finding using full granularity, resolution**
- **Then: apply full track reconstruction, match to ECAL cluster**

(work of E. Meschi)

HLT: electrons (II)



HLT: muons (I)

Problem: most muons are real

- **They may be non-prompt (e.g. K→$\mu\nu$) but they are real**
- **Need to generate full events, that can be simulated afterwards, containing the correct mix of muon origins (e.g. heavy flavors, decays, J/ψ→$\mu\mu$, etc...)**
- **Solution:**

  (a) for each event generated by PYTHIA, look at set of final-state pseudo-scalar mesons (B, D, π, K, ρ, ...)

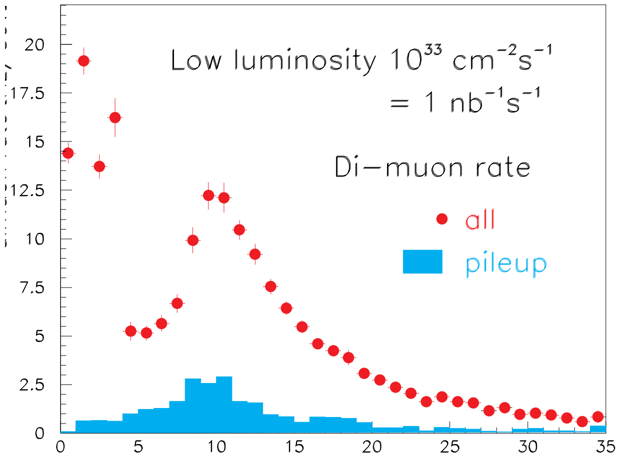
  (b) compute probability that a muon (passing acceptance requirements) will appear. "Force" this configuration, store event weight.

  (c) Then simulate event, Lvl-1 μ trigger, apply cuts...
- **Method extendable to any μ multiplicity (e.g. dimuons)**

(work of H. Rick)

HLT: muons (II)

Example:
dimuon rate



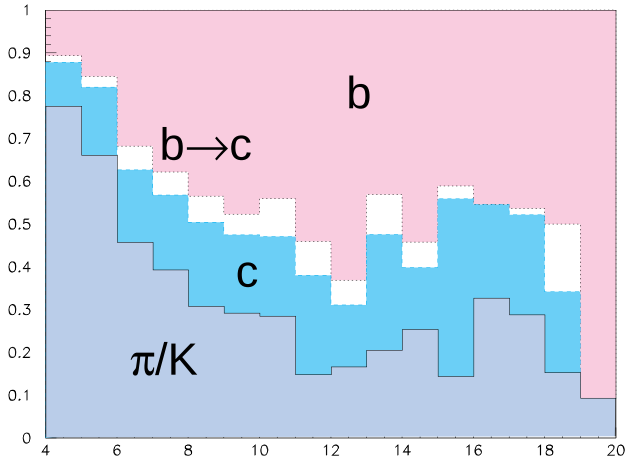
Muon sources

Decayed parents of muons at $p_T^\mu > 6\,\mathrm{GeV}$:

	$B^\pm$	B^0	B^0_s	Λ^0_b	$\Xi^{\pm 0}_b$	$\pi^\pm$	$K^\pm$	K^0_L
%	16.3	14.4	5.3	3.2	0.4	20.3	17.6	0.2
%	39.6					38.1		

	D^0	$D^\pm$	$D^\pm_s$	J/ψ	$\Lambda^\pm_c$	$\tau^\pm$	η	ρ^0
%	10.5	6.8	2.7	0.9	0.4	0.8	0.1	0.1
%	21.3					1.0		

Heaviest parton in decay history (rel. contributions):



HLT: tracking

Issue: how much tracking info needed?

- **Regional algorithms: given seed (e.g. e, μ at Lvl-1) determine road in tracker that should contain hits from the (particle's) track. Ongoing work (MSGCs ok)**

  (T. Monteiro)

- **Once road is identified, call on track reconstruction to find tracks using only detector modules in road**

  (S. Khanov, N. Stepanov)

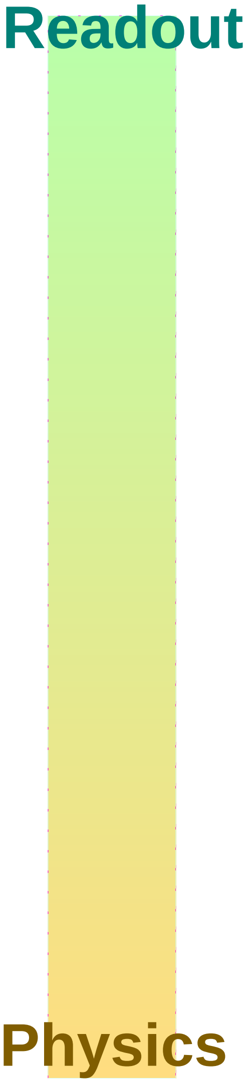
- **Preliminary results:**

  (a) For electrons one should start from inner layers (e.g. pixels); tracker material → lots of radiation, so MSGC stubs don't help much

  (b) For muons, starting from outer detector layers (MSGCs) should be ok

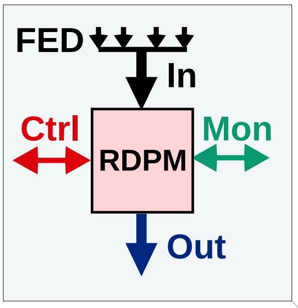
HLT: Ultimate Goals

Issues for HLT:



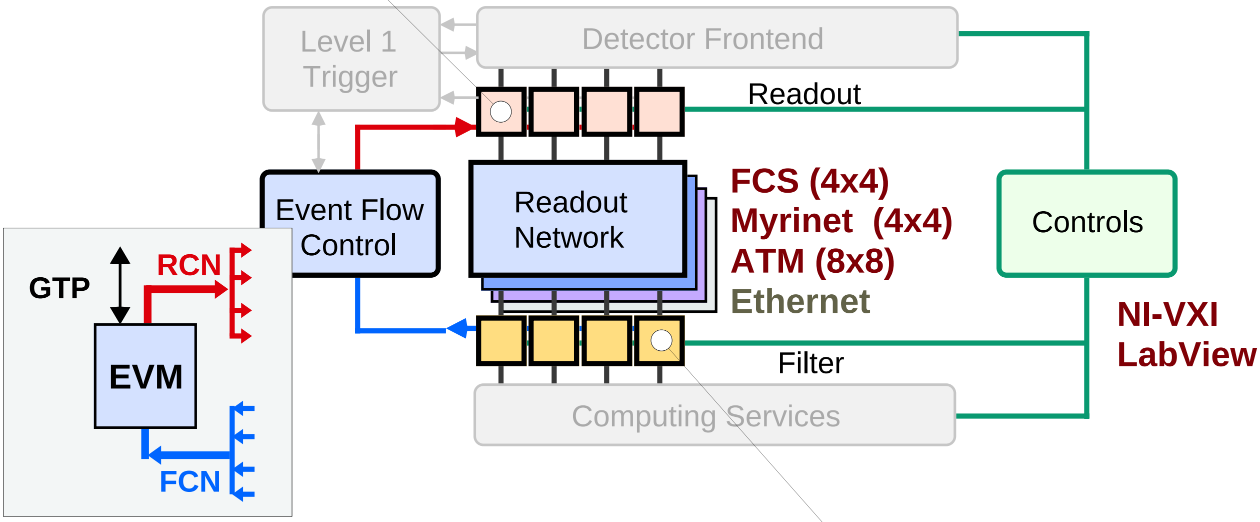
- Basic Unit of Information (parton model) of CMS readout
- Tradeoff(s) between small data access and efficiency of the data transfer: small blocksizes → low efficiency
- Depending on link utilization efficiency, may have a pre-Event Building step → need a hadron model of CMS readout
- Implementation of Level-2 algorithms; resulting rates
- How many trigger levels? (Could have continuum...)
- Amount of information needed by Levels 2 and 3
- Selection criteria for what ends up on tape

DAQ 98 demonstrators summary



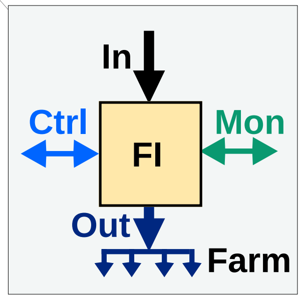
RDPM:	FPGA, PPC (VxWorks)
Input :	Pattern generator
Ctrl:	Ethernet, (FCS, Myrinet)
Output:	ATM, FCS, Myrinet
Mon:	NI-VXI (LabVIEW)

10..100 kHz I/O logics
60..100 MB/s memory I/O



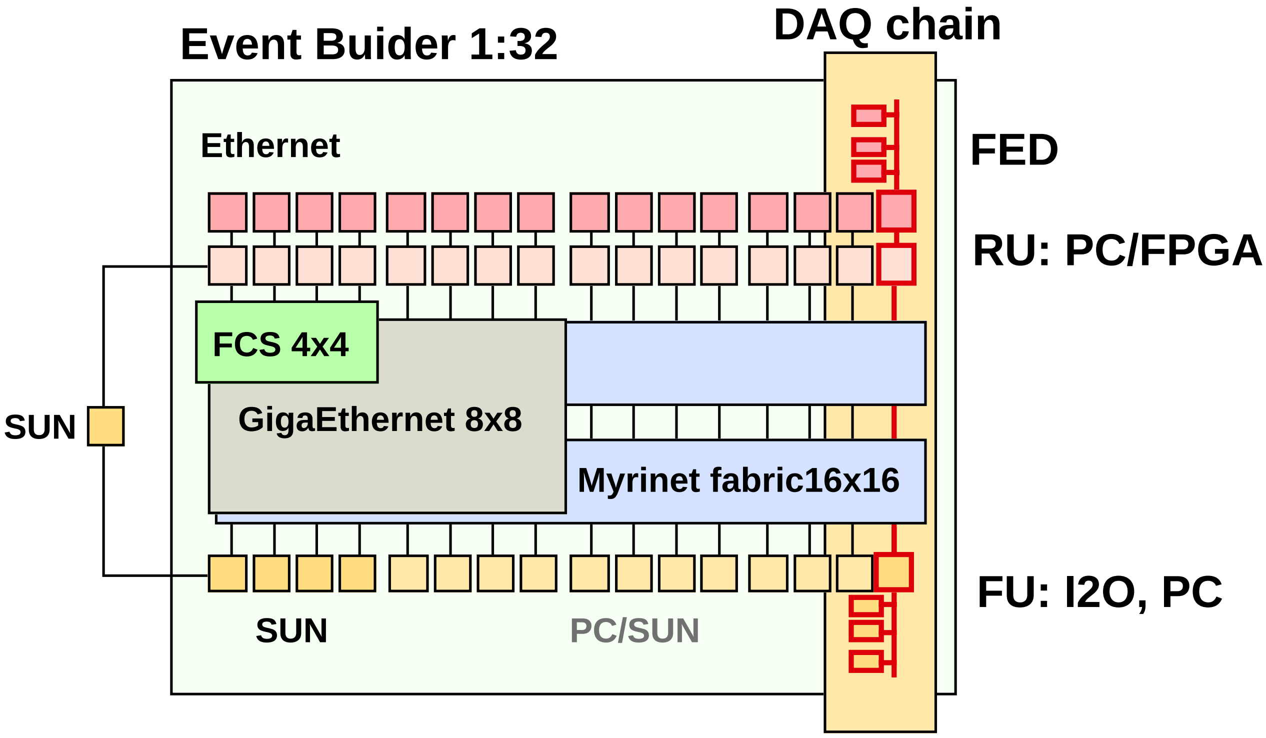
EVM:	SUN (Solaris)
GTP :	Internal
RCN:	Ethernet, (FCS, Myrinet)
FCN:	Ethernet, (FCS, Myrinet)

40 kHz messages



FI:	SUN (Solaris), PPC, C80
Input :	FCS, Myrinet, ATM
Ctrl:	Ethernet
Output:	Dummy, OODB
Mon:	Ethernet

DAQ 99 demonstrators



DAQ chains in operation:

Studies of the parameters of the CMS DAQ design via simulation and test benches.

RU prototype 2 (design: FPGA & PC)
EVB Prototype 1 complete (Myrinet, GE)

RU prototype 2 complete
Vertical DAQ chain prototype (FED-RU-FU)
HLT prototype 1

May 99 milestones

DAQ chains in operation:

Studies of the parameters of the CMS DAQ design via simulation and test benches.

RU prototype 2 (design): The RU prototype will provide the full functionality required for the DAQ demonstrator of the year 2000.
- **RUM:** design of a dual-ported intelligent memory module (RDPM-P3) with two PCI buses (32/64 bit, 33/66 MHz) addressable via a high-level protocol.
- **Software:** This is the software that will run under VxWorks in the RUI and RUO CPUs, suitable for use in both the EVB test-bench and in small DAQ systems.

EVB Prototype 1 complete: Operation of a **full Event Builder with 16 ports,** some of which will be connected to full RU prototypes and FU prototypes. The system will be supplemented with a corresponding simulation. The EVB prototype (and its associated simulation) will also be used to evaluate the effectiveness and applicability of various traffic-shaping and event-building protocols.
- **EVB prototype 8x8:** This EVB prototype will be an extension of the 4x4 system of 1998 to 8x8. It will consist of 8x8 switches (of multiple technologies) in order to study performance and fabric structures.
- **Control Network Prototype 1:** The EVB will utilize a first prototype of the Control Networks (RCN and FCN), most likely based on Ethernet.

November 99 milestones

RU prototype 2 complete: Complete tests of RU prototype 2.

FU prototype 1 :

- Filter Unit Prototype 1: The FU prototype will be implemented using a commercial CPU (either a workstation or a PC) supplemented with a Network Interface card and/or a FI card. This CPU will be connected via Fast Ethernet to a cluster of PCs/workstations running a dummy HLT filtering process.

- Hardware for FU P1: FI card, commercial CPU, sub-farm, Ethernet network

- Software for FU P1: FU Control Software running on the commercial CPU, FI driver, FI switch software, sub-farm data-flow software, prototype HLT module.

Vertical DAQ chain prototype: Integration and operation of a complete DAQ chain including:

- Detector front-end (FED),

- Readout Unit

- Data network (between RU and FU)

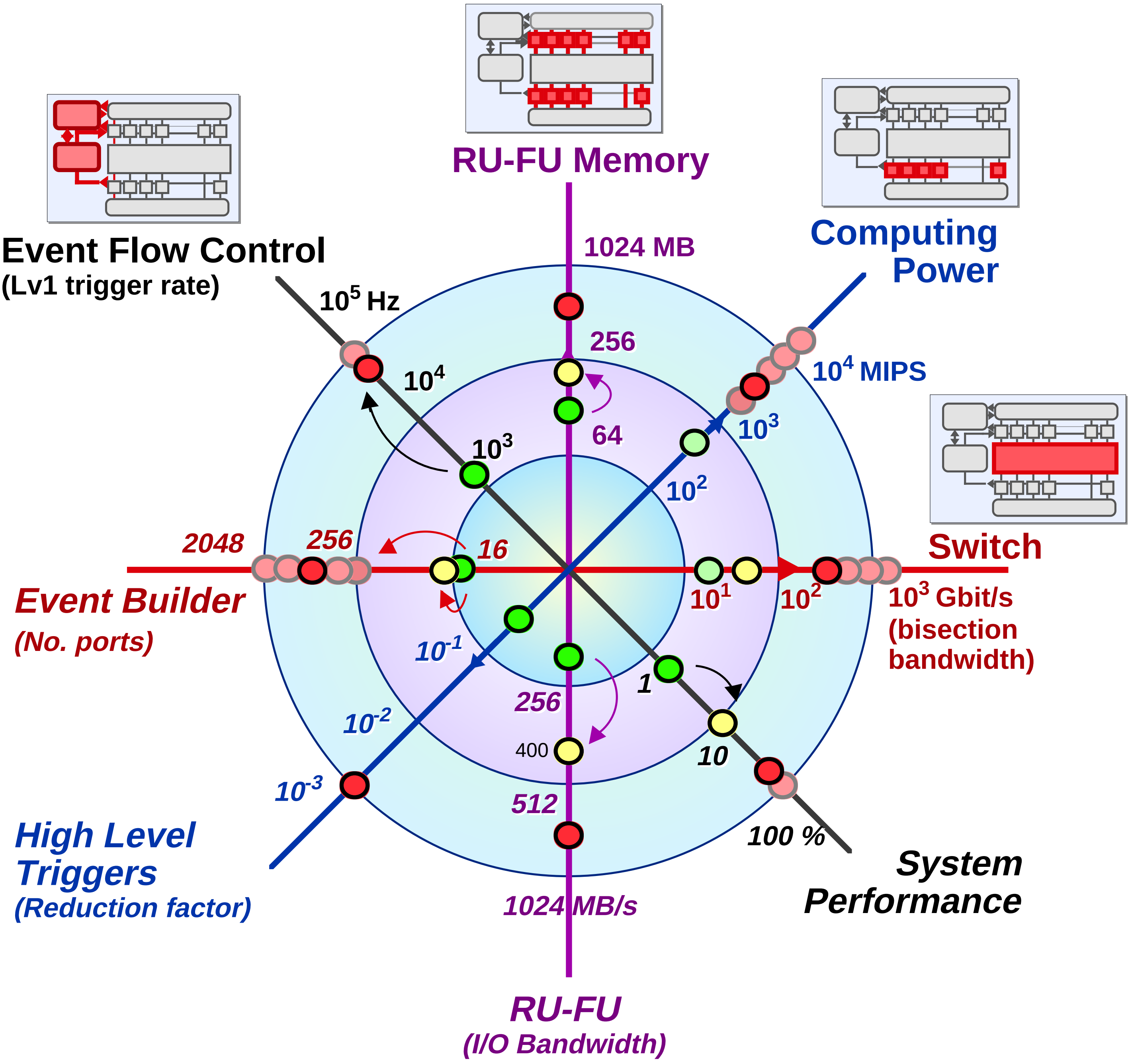
- Filter Unit. These will be used both in test beams and laboratory DAQ demonstrators.

HLT prototype 1

- Results on rates and efficiencies for all Level-1 Triggers

- Prototype of steering software in C++

DAQ parameters & 99 prototypes



RU-FU Memory

Event Flow Control
(Lv1 trigger rate)

Computing
Power

Switch

Event Builder
(No. ports)

High Level
Triggers
(Reduction factor)

System
Performance

RU-FU
(I/O Bandwidth)

Cost book, organization, calendar

Cost book 9 partitions

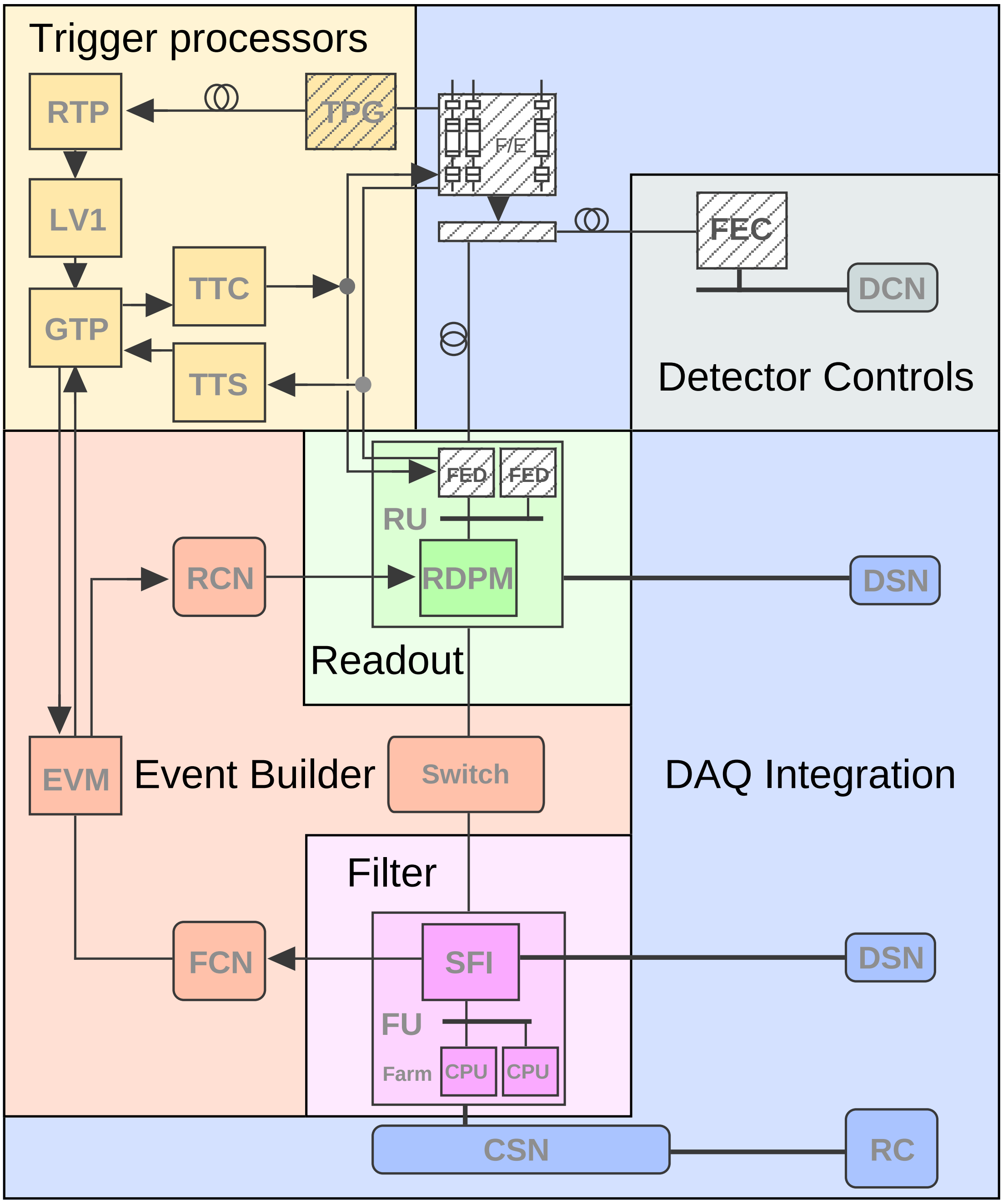
TriDAS institutions and sharing of resposabilities

TriDAS organization

TriDAS bodies

TriDAS 99 events and URLs

Cost book 9 partitions



Cost book 9. DAQ L4 entries

DAQ

Readout Unit

RUI	Readout Unit Input
RUM	Readout Unit Memory
RUO	Readout Unit Output
RUS	RU Supervisor
Readout Unit Software	
Crates	
Prototypes	

Filter Unit

FUI	Filter Unit Input
FUM	Filter Unit Memory
FUO	Filter Unit Output
FUS	FU Supervisor
Filter Unit Software	
Crates	
Filter Node farm	
Prototypes	

Event Builder

EVM	Event Manager
RCN	Readout Control Network
FCN	Farm Control Network
Switch	
Prototypes	

DAQ Integration

CSN	Computing and Services Network
DSN	DAQ Services Network
Farm control	
Control room	
Mass storage	
DAQ Software	

TriDAS cost book 9

	CMS Cost Estimate - Version 9		Release Date: March 31, 1998			
No.	Item	Unit Type	# Units	Unit Cost	Total Cost	
6.1	**TRIGGER**				12140	
6.1.1	CALORIMETER TRIGGER				5225	
6.1.2	CSC TRIGGER				1100	
6.1.3	DT TRIGGER				780	
6.1.4	RPC TRIGGER				3695	
6.1.5	GLOBAL TRIGGER				1340	
6.2	**DATA ACQUISITION**				23043	
6.2.1	READOUT UNIT				5559	
6.2.2	FILTER UNIT				10638	
6.2.3	EVENT BUILDER				5318	
6.2.4	DAQ INTEGRATION				1528	
6.3	**DETECTOR CONTROLS**				2342	
6.3.1	DETECTOR CONTROLS				2342	
	Total: Trigger / DAQ				37525	

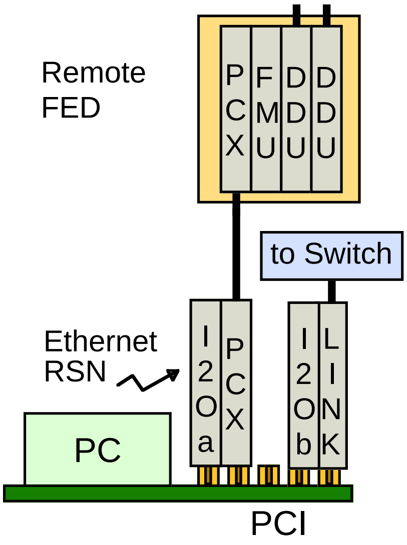
Trigger cost book 9

	CMS Cost Estimate - Version 9		**Release Date: March 31, 1998**			
No.	Item	Unit Type	# Units	Unit Cost	Total Cost	
6.1.1	**CALORIMETER TRIGGER**				5225	0
6.1.1.1	Regional trigger	System	1	4050.020	4050	
6.1.1.2	Global Cal. Trigger	System	1	400.000	400	
6.1.1.3	Readout & Control	System	1	255.000	255	
6.1.1.4	Data Communication	System	1	520.000	520	
6.1.2	**CSC TRIGGER**				1100	0
6.1.2.1	Muon Port Card	Board	63	7.475	471	
6.1.2.2	Sector Receivers	Board	65	4.217	274	
6.1.2.3	Sector Processors	Board	15	7.020	102	
6.1.2.4	Overlap Processors	Board	15	7.020	105	
6.1.2.5	Clock and Control Card	Board	12	5.233	63	
6.1.2.6	Crate Monitor Card	Board	10	1.300	13	
6.1.2.7	Trigger Crate	Crate	10	7.107	71	
6.1.2.7	Institute Manpower	Manyears	0	0.000	0	
6.1.3	**DT TRIGGER**				780	0
6.1.3.1	Trigger Crate	Crate	4	6.000	24	
6.1.3.2	Sector Processor Card	Board	60	8.000	480	
6.1.3.3	Muon Sorter Card	Board	5	8.000	40	
6.1.3.4	Clock and Control Card	Board	4	4.000	16	
6.1.3.5	Crate Monitor Card	Board	4	2.000	8	
6.1.3.6	Cables	Cables	1	50.000	50	
6.1.3.7	Readout and Control	System	1	32.000	32	
6.1.3.8	Monitoring and Test	System	1	70.000	70	
6.1.3.9	Prototypes and spares	System	1	60.000	60	
6.1.3.10	Institute Manpower	Manyears	0	0.000	0	0
6.1.4	**RPC TRIGGER**				3695	0
6.1.4.1	Link Board	Board	936	0.475	445	
6.1.4.2	Data Communication	System	1	755.160	755	
6.1.4.3	Trigger Board (Barrel)	Board	156	3.830	597	
6.1.4.4	Trigger Board (EndCap)	Board	240	3.350	804	
6.1.4.5	Final Sorter Board	Board	33	1.600	53	
6.1.4.6	Readout board	Board	396	0.700	277	
6.1.4.7	Readout Concentrator Board	Board	33	1.000	33	
6.1.4.8	Clock Control Board	Board	33	4.000	132	
6.1.4.9	Trigger Crate	Crate	33	8.600	284	
6.1.4.10	Readout and Control	System	1	15.300	15	
6.1.4.11	ASICs Development	Service	1	200.000	200	
6.1.4.12	Prototypes and spares	System	1	100.000	100	
6.1.4.13	Institute Manpower	Manyears	0	0.000	0	0
6.1.5	**GLOBAL TRIGGER**				1340	0
6.1.5.1	Global Trigger Crate	Crate	1	27.046	27	
6.1.5.2	PipelineSyncBuffer	Board	5	8.290	41	
6.1.5.3	Global Trigger Logic	Board	5	12.168	61	
6.1.5.4	Global Trigger Final	Board	1	5.199	5	
6.1.5.5	Cables	Cables	1	4.070	4	
6.1.5.6	Global Muon Trigger	System	1	158.272	158	
6.1.5.7	Readout and Control	System	1	30.000	30	
6.1.5.8	Prototypes	System	1	32.700	33	
6.1.5.9	Monitoring & Test	System	1	80.000	80	
6.1.5.10	Trigger Throttle System	System	6000	0.150	900	
6.1.5.11	Institute Manpower	Manyears	0	0.000	0	0

DAQ cost book 9

	CMS Cost Estimate - Version 9		Release Date: March 31, 1998			
No.	Item	Unit Type	# Units	Unit Cost	Total Cost	
6.2.1	**READOUT UNIT**				5559	
6.2.1.1	Readout Unit Input (RUI)	Board	563	1.700	957	
6.2.1.2	Readout Unit Memory (RUM)	Board	563	3.000	1689	
6.2.1.3	Readout Unit Output (RUO)	Board	436	2.500	1090	
6.2.1.4	Readout Unit Supervisor (RUS)	Board	436	1.750	763	
6.2.1.5	Crates	Crate	436	1.000	436	
6.2.1.6	RU Software	Licence	436	0.700	305	
6.2.1.7	Prototypes	System	1	318.400	318	
6.2.1.8	Institute Manpower	Manyears	250			
6.2.2	**FILTER UNIT**				10638	
6.2.2.1	Filter Unit Input (FUI)	Board	436	2.500	1090	
6.2.2.2	Filter Unit Memory (FUM)	Board	436	2.000	872	
6.2.2.3	Filter Unit Output (FUO)	Board	436	1.700	741	
6.2.2.4	Filter Unit Supervisor (FUS)	Board	436	1.750	763	
6.2.2.5	Crates	Crate	436	1.000	436	
6.2.2.6	Filter Node Farm	CPU	512	11.300	5786	
6.2.2.7	FU Software	Licence	384	0.788	303	
6.2.2.8	Prototypes	System	1	648.000	648	
6.2.2.9	Institute Manpower	Manyears	200			
6.2.3	**EVENT BUILDER**				5318	
6.2.3.1	Event manager (EVM)	System	1	150.000	150	
6.2.3.2	Readout Control Network (RCN)	System	1	102.400	102	
6.2.3.3	Farm Control Network (FCN)	System	1	204.800	205	
6.2.3.4	Switch	System	1	4608.000	4608	
6.2.3.5	EVB Software	Licence	0	0.000		
6.2.3.6	Prototypes	System	1	253.260	253	
6.2.3.7	Institute Manpower	Manyears	40			
6.2.4	**DAQ INTEGRATION**				1528	0
6.2.4.1	Computing&Services Network (CSN)	System	512	0.200	102	
6.2.4.2	DAQ Services Network (DSN)	System	1126	0.200	225	
6.2.4.3	Farm Control	System	0	0.000	0	
6.2.4.4	Mass Storage	System	10	50.000	500	
6.2.4.5	Control Room	System	500	1.000	500	
6.2.4.6	DAQ Software	Licence	200	1.000	200	
6.2.4.7	Institute Manpower	Manyears	0			
6.3.1	**DETECTOR CONTROLS**				2342	
6.3.1.1	DCS Equipment	System	1	1936.000	1936	
6.3.1.2	DCS Test Facilities	System	1	212.000	212	
6.3.1.3	Prototypes	System	1	194.000	194	
6.3.1.4	Institute Manpower	Manyears	30		0	

RU-FU costing example

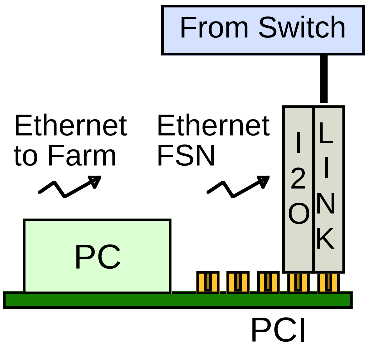


Desktop RU. Description

RU system based on desktop PC. The CPU runs the RDPM event handling logics, the data buffering and the unit monitoring (RUS, RUM). The FED system is readout via a simple PCI-to-PCI connection (PCX) driven by an I2O board (RUI), a second I2O (RUO) drives the output link reading data from the PC memory via PCI. EVM commands (RUIC/RUOC) are distributed via ethernet port in I2Oa(RUI) and I2Ob(RUO).

Equivalance		**Cost (CHF)**
RUI, RUI bus	I2Oa	1500 (i960 IOP)
	PCI-to-PCI (PCX)	1000 (e.g. NI-MXI)
RUS, RUM, DSN	Desktop CPU, PC-RAM, ethernet, PCI	5000 (sun PCI)
Crate, RUO bus		
RUO	I2Ob	1500 (i960 IOP)
	PCI link interface	1000 (ATM)

Total 10000



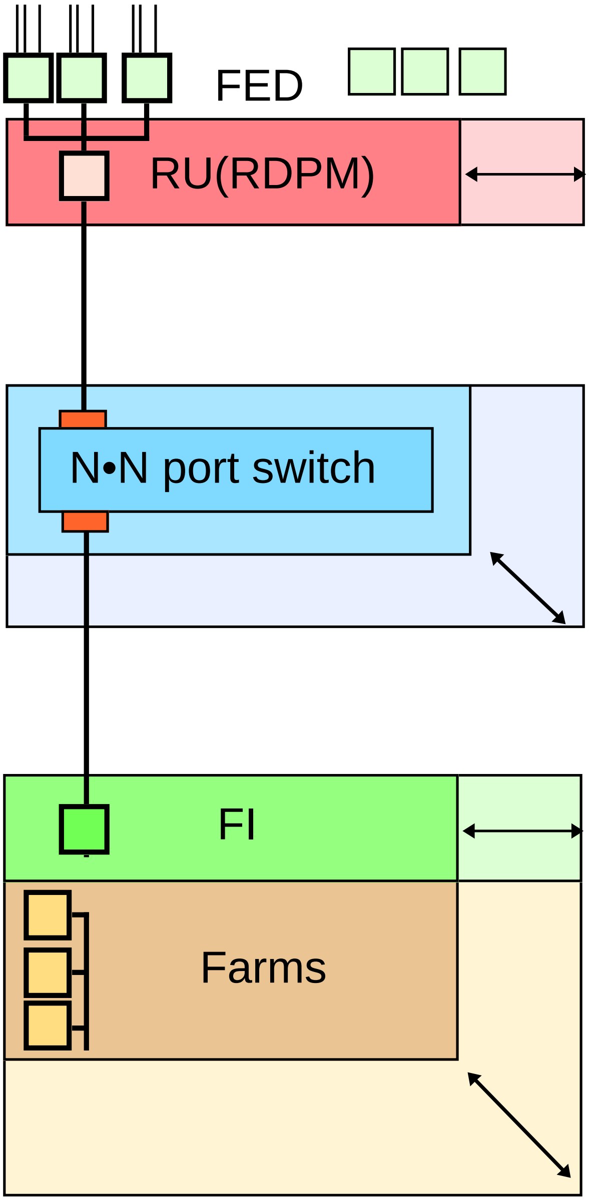
Desktop FU. Description

FU system based on desktop PC. The event data fragment input link is driven by an I2O module. The CPU runs the event data memory handling logics, the data buffering, the unit monitoring and the communication with the farm (FUS, FUM, FUO). The connection with the processor farm is implemented by ethernet (or a PCI-to-PCI extension).

Equivalance		**Cost (CHF)**
FUI, FSN	I2O, I2O-ethernet	1500 (i960 IOP)
FUI bus,	PCI link interface	1000 (ATMI)
FUS, RFUM, DSN	Desktop CPU, PC-RAM, ethernet	5000 (sun PCI)
FUO, Crate		
FUO bus	PCI-ethernet	500

Total 8000

Data acquisition scaling



Readout Unit (FED+RDPM)
FrontEnd Driver (FED) : **NOSCALING** at this stage (scaling the number of FED means reducing the detector number of channels
RDPM : the number of memories and switch ports can be scaled with the switch size
(Contingency:Event size*trigger rate)

Event Builder: the switch is scalable as:
- Number of ports
- Port speed
- Switch internal resources

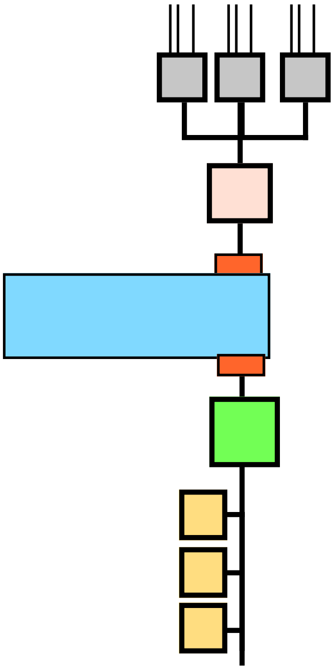
(Contingency:Event size*trigger rate+Switch load)

Filter Unit (FI+FN)
The number of FIs is scalable with switch size
Filter Nodes (FN) are scalable:
- As number of CPUs
- Very likely the farm will be made of power servers based on internal multiprocessor (scalable) architecture

(Contingency:Levl2- rejection, HLT algorithms)

Cost Book 9
- Frontend readout at 100 kHz
- Event builder and Farms scaled to 75 kHz

DAQ readout rate and costs



DAQ column cost:

RDPM	10.000
Switch ports	9.000
FI	8.000
Filter Node	11.000
TOTAL	**38.000 CHF**

A DAQ column (in the current baseline design in CB9)contributes to 1/500 of the maximum level-1 trigger rate that is 200 Hz

The cost to readout and process one level-1 trigger Hz is about 200 CHF

TriDAS institutions and sharing of resposabilities

Country	Institute	Trigger Calo.	Trigger Muon	Trigger Global	Readout	Event Builder	Event Filter	Software
Austria	Vienna		▨	▨				
CERN				▨	▨	▨	▨	▨
Finland	HTI Helsinki	▨			▨			
France	Saclay				▨		▨	
	LPNHE	▨					▨	
Greece	Democritos				▨			
	Ioannina				▨			
	Crete				▨			
Italy	Bari		▨					
	Bologna		▨					
	Padova		▨					
	Pisa		▨					
Poland	Warsaw		▨					
Portugal	Lisbon	▨						▨
Switzerl.	EPFL					▨		
	ETH				▨	▨		
	PSI				▨			
Hungary	KFKI-RMKI				▨			
UK	Bristol			▨				
	Imp. College				▨			
	RAL				▨			▨
USA	Fermilab	▨			▨	▨	▨	▨
	Iowa State					▨		
	Nebraska	▨						
	Missisipi					▨		
	MIT				▨	▨	▨	▨
	Rice		▨					
	UC Davis		▨					
	UCLA		▨				▨	
	UCSD				▨	▨	▨	
	Wisconsin	▨						

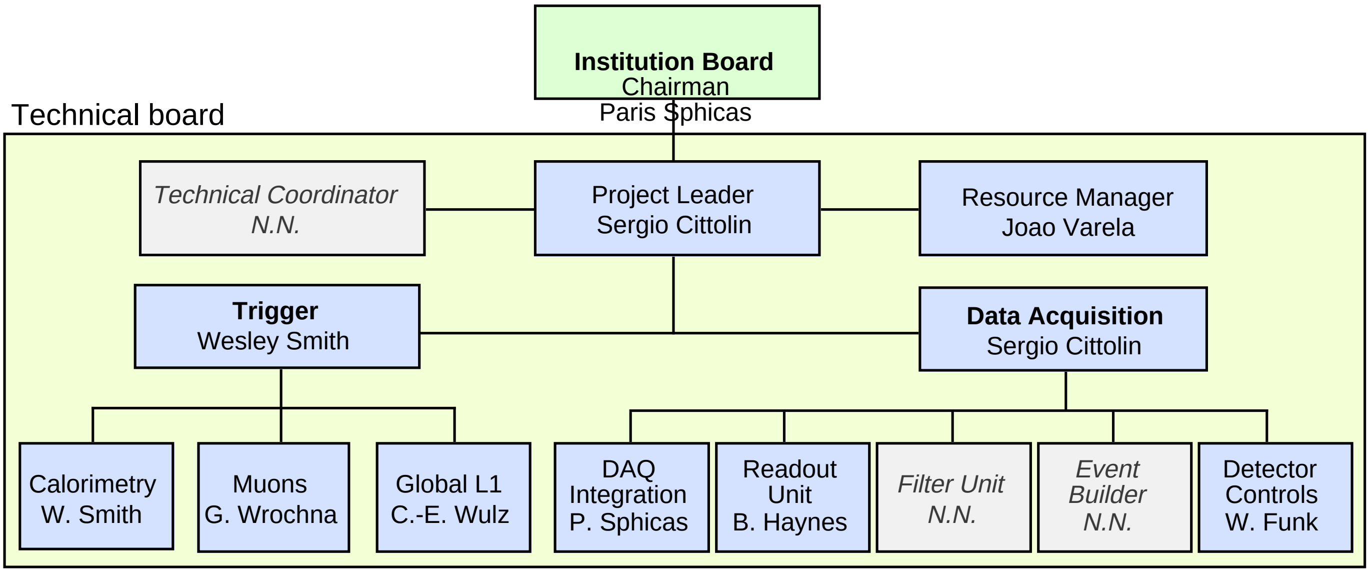
TriDAS Money Matrix

TriDAS Money Matrix V2.1

March 31, 1998

No.	Item	Total Cost V9	AUSTRIA	CERN	FINLAND	FRA Cea	GREECE	HUNGARY	ITALY	KOREA	POLAND	PORTUGAL	SWI Eth	SWI Psi	U.K.	USA Doe	USA Nsf	**Assigned**	**Balance**
6.1	**TRIGGER**	12140	1220	200	1020	0	700	0	100	1000	2060	255	0	0	400	5150	0	12105	-35
6.1.1	CALORIMETER TRIGGER	5225			520							255			400	4050		5225	0
6.1.2	CSC TRIGGER	1100														1100		1100	0
6.1.3	DT TRIGGER	780	780															780	0
6.1.4	RPC TRIGGER	3695			500				100	1000	2060							3660	-35
6.1.5	GLOBAL TRIGGER	1340	440	200			700											1340	0
6.2	**DATA ACQUISITION**	23043	30	10425	0	840	1360	90	0	0	0	0	5505	500	450	3599	763	23562	518
6.2.1	READOUT UNIT	5559		2120			1360						1389	350	450			5669	110
6.2.2	FILTER UNIT	10638	30	2775		840		90					2872	150		3426	763	10946	308
6.2.3	EVENT BUILDER	5318		4000									1244			173		5417	98
6.2.4	DAQ INTEGRATION	1528		1530														1530	2
6.3	**DETECTOR CONTROLS**	2342	0	2345	0	0	0	0	0	0	0	0	0	0	0	0	0	2345	3
6.3.1	DETECTOR CONTROLS	2342		2345														2345	3
		37525																	
	Assigned		1250	12970	1020	840	2060	90	100	1000	2060	255	5505	500	850	8749	763	38011	
	Expected Contributions		1250	12970	1020	840	2060	90	100	1000	2060	255	5505	500	850	8749	763	38012	487
	Balance		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

TriDAS organization



Institution Board
Chairman
Paris Sphicas

Technical board

Technical Coordinator
N.N.

Project Leader
Sergio Cittolin

Resource Manager
Joao Varela

Trigger
Wesley Smith

Data Acquisition
Sergio Cittolin

Calorimetry
W. Smith

Muons
G. Wrochna

Global L1
C.-E. Wulz

DAQ
Integration
P. Sphicas

Readout
Unit
B. Haynes

Filter Unit
N.N.

Event
Builder
N.N.

Detector
Controls
W. Funk

TriDAS	S. Cittolin
Level1 Trigger	W. Smith
Calorimeter	W. Smith
Trigger Primitives	P. Busson
Regional Trigger	W. Smith
Global Cal. Trigger	G. Heath
Readout and Control	S. Silva
Muon	G. Wrochna
RPC	J. Krolikowski, M. Kudla
CSC	J. Hauser, T.Y. Ling
DT	P. Zotto, R. Martinelli
Track Finder	F. Szoncso
Global Muon Trigger	A. Taurok
Global Trigger	C. Wulz
Processor	
Monitoring	
Data Acquisition	S. Cittolin
DAQ Integration	P. Sphicas
Hardware prototypes	A. Racz
SFI&FilterNode	J. Branson
High Level Triggers	P. Sphicas
DAQ Software	D. Samyn
Readout Unit	B. Haynes
Filter Unit	NN
Event Builder	NN
Detector Controls	W. Funk
Project support	
Resource Manager	J. Varela
Plan	P. Sphicas, W. Smith

TriDAS bodies

TriDAS Institution Board

AUSTRIA	HEPHY - Wien	Claudia, Wulz
FINLAND	Univ Helsinki	Tuominiemi, Jorma
	HIP - Helsinki	Tuominiemi, Jorma
	Tampere Univ	Tuominiemi, Jorma
FRANCE	CEA-Saclay	Faure, Jean-Louis
GREECE	Demokritos	Fanourakis, George
	Univ of Ioannina	Manthos, Nikos
HUNGARY	KFKI Budapest	Vesztergombi, Gyorgy
ITALY	Bari	Iaselli, Giuseppe
	Bologna	DallaValle, Marco
KOREA		
POLAND	IEP,Warsaw	Krolikowski, Jan
	SINS, Warsaw	Krolikowski, Jan
PORTUGAL	LIP Lisbon	Varela, Joao
SWITZERLAND	CERN	Cittolin, Sergio
	ETH	Hofer, Hans (Rubbia, Andre)
	PSI	Kotlinski, Danek
UK	Bristol Univ	Heath, Gregory
	RAL	Haynes, Bill
USA	UC Davis	Ko, Winston
	UC Los Angeles	Hauser, Jay
	UC San Diego	Branson, Jim
	Carnegie Mellon Univ	
	FNAL	Gaines, Irwin
	Univ Florida, Gainesville	Acosta, Darin
	Iowa State Univ	Hauptman, John
	MIT	Sphicas, Paris, IB Chairman
	Ohio State Univ	Ling, T.Y.
	Rice Univ	Padley, Paul
	Univ Wisconsin	Smith, Wesley

TriDAS technical board

US IB Chairman	J. Branson
DAQ	S. Cittolin, TB Chairman
Readout Unit	B. Haynes
Data Links	E. Pietarinen
Detector Controls	W. Funk
Level-1 Trigger	W. Smith
DAQ Integration	P. Sphicas
Resource manager	J. Varela
Muon Trigger	G. Wrochna
CalorimeterTrigger	W. Smith
Global Trigger	C. Wulz

FERU working group

Chair	B. Haynes
Secretary	A. Racz
Tracker	R. Halsall
Pixel	D. Kotlinski
HCAL	J. Elias
ECAL	M. Hansen
Preshower	S. Reynault
RPC	M. Kudla
DT	J. Marin
CSC	T.Y. Ling, S. Durkin
DCS	W. Funk

+ S. Cittolin, W.Smith, P. Sphicas

Participation to LHCC common projects

- Detector Control Systems

- Filter Unit

TriDAS 99 events and URLs

17	March	Calibration WS
17	May	CMS TriDAS referees
17-21	May	TriDAS internal review #1
16	June	FE/RU&Sync II WS
27	October	High Level Triggers WS
28	October	Detector Controls WS
8	November	CMS TriDAS referees
8-12	November	TriDAS internal review #2

TriDAS REFEREE Web page :

http://cmsdoc.cern.ch/ftp/afscms/TRIDAS/html/Referee.html

See also TriDAS home page at:

http://cmsdoc.cern.ch/ftp/afscms/TRIDAS/html/tridas.html