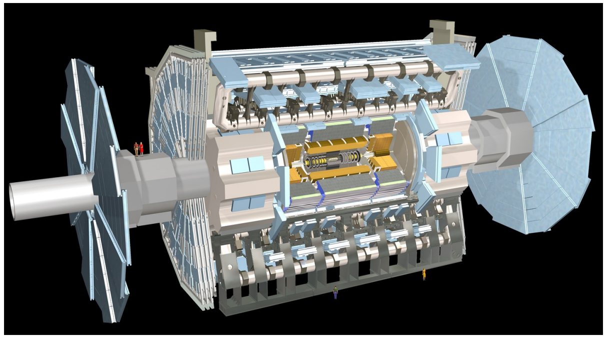
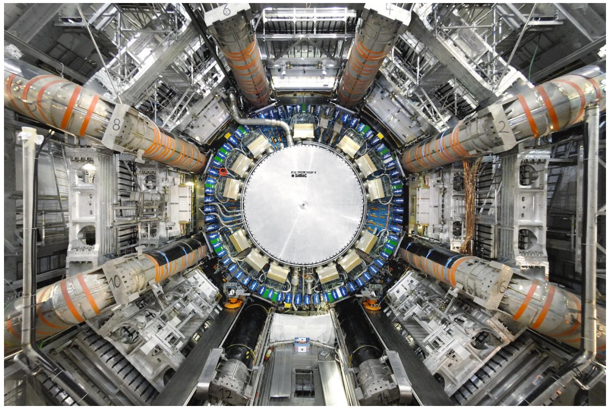




# ATCA Test Platform (RCE/CIM Development Lab)

*ROD Workshop*

*19 June, 2009*

**Rainer Bartoldus**
**SLAC**

# Outline

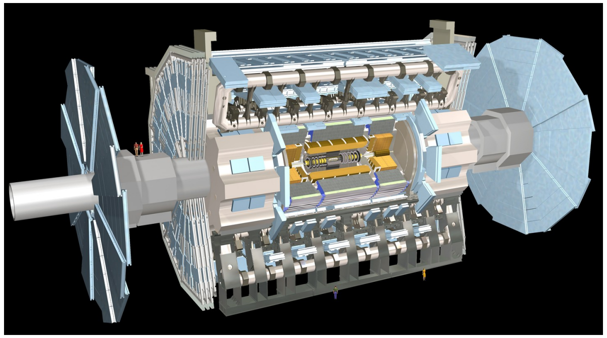
- **Generic DAQ Building Blocks**
  - Reconfigurable Cluster Element (RCE)
  - Cluster Interconnect (CI)
  - Substrate ATCA (Covered by Markus in previous talk)
- **The RCE/CIM Test Platform**
  - Current Installation
  - RCE Training Workshop
  - Where to go from here
- **Looking Ahead (or in Philippe's words: “Wild Imagining”)**
  - A Hypothetical 48-Channel Read Out Module
  - Bandwidth Calculations on a Napkin
- **Summary**

# More Information

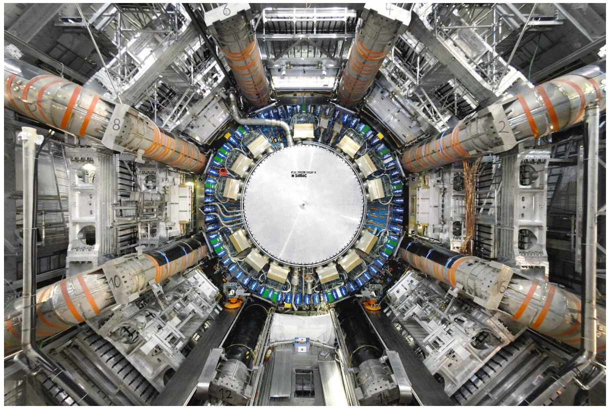
- **RCE Training Workshop Page**
  - `http://indico.cern.ch/conferenceDisplay.py?confId=57836`
  - Links to workshop presentations on e.g. (cross-)development cycle, RTEMS, class libraries and APIs, code examples

- **RCE Lab TWiki**
  - `https://twiki.cern.ch/twiki/bin/view/Atlas/RCEDevelopmentLab`
  - Description and setup of the RCE lab infrastructure, host names, instructions, external material, e.g., ATCA manuals etc.

- **RCE High Lumi Mailing List**
  - `https://groups.cern.ch/group/atlas-highlumi-RCE-development`
  - Discussions on Phase II upgrade studies using the RCE platform, announcements of future workshops



# Building Blocks



# Three Building Block Concepts

- **Computational Elements**
  - Must be low-cost
    - $$$, footprint, power
  - Must support variety of computational models
  - Must have both flexible and performant I/O

**The Reconfigurable Cluster Element (RCE) based on:**
- **System-On-Chip technology**
- ***Virtex* 4 & 5**

- **Mechanism to Connect Together these Elements**
  - Must be low-cost
  - Must provide low-latency/high bandwidth I/O
  - Must be based on commodity (industry) protocol
  - Must support a variety of interconnect topologies
    - Hierarchical, peer-to-peer, fan-in & fan-out

**The Cluster Interconnect (CI) based on:**
- **10 Gb Ethernet switching**

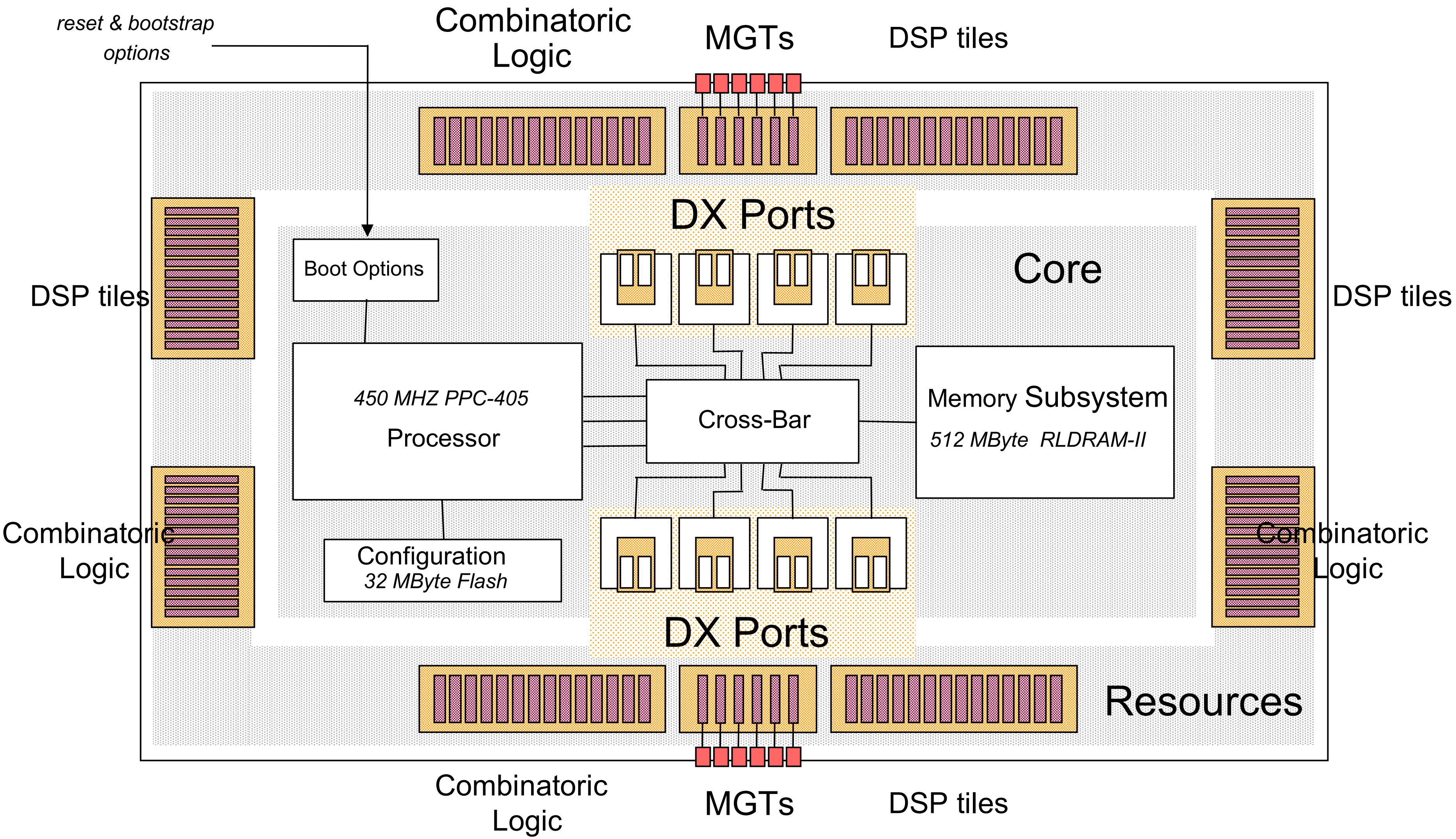
- **Packaging Solution for Both Element and Interconnect**
  - Must provide high availability
  - Must allow scaling
  - Must support different physical I/O interfaces
  - Preferably based on a commercial standard

**ATCA:**
- **Crate based**
- **Serial backplane**

# A Word on ATCA...

- ***One* logical choice for a new packaging standard**
  - There are other possibilities, c.f. Markus' talk
  - It has very attractive features
    - e.g. Rear Transition Module (RTM) and high-speed serial backplane, protocol-agnostic, providing different topologies
  - People who have worked with it tend to like it a lot
- **It is still only a substrate for the other two building blocks**
  - Albeit a rather ideal one
- **The RCE/CIM concept can be mounted on other standards that provide similar benefits**
  - One compelling, non-technical reason for picking ATCA now is that you can actually *buy* a crate (shelf) today
  - (BTW, as ATCA becomes increasingly popular, it no longer makes too much sense to refer to the RCE/CIM platform as the ATCA platform...)

# (Reconfigurable) Cluster Element (RCE)



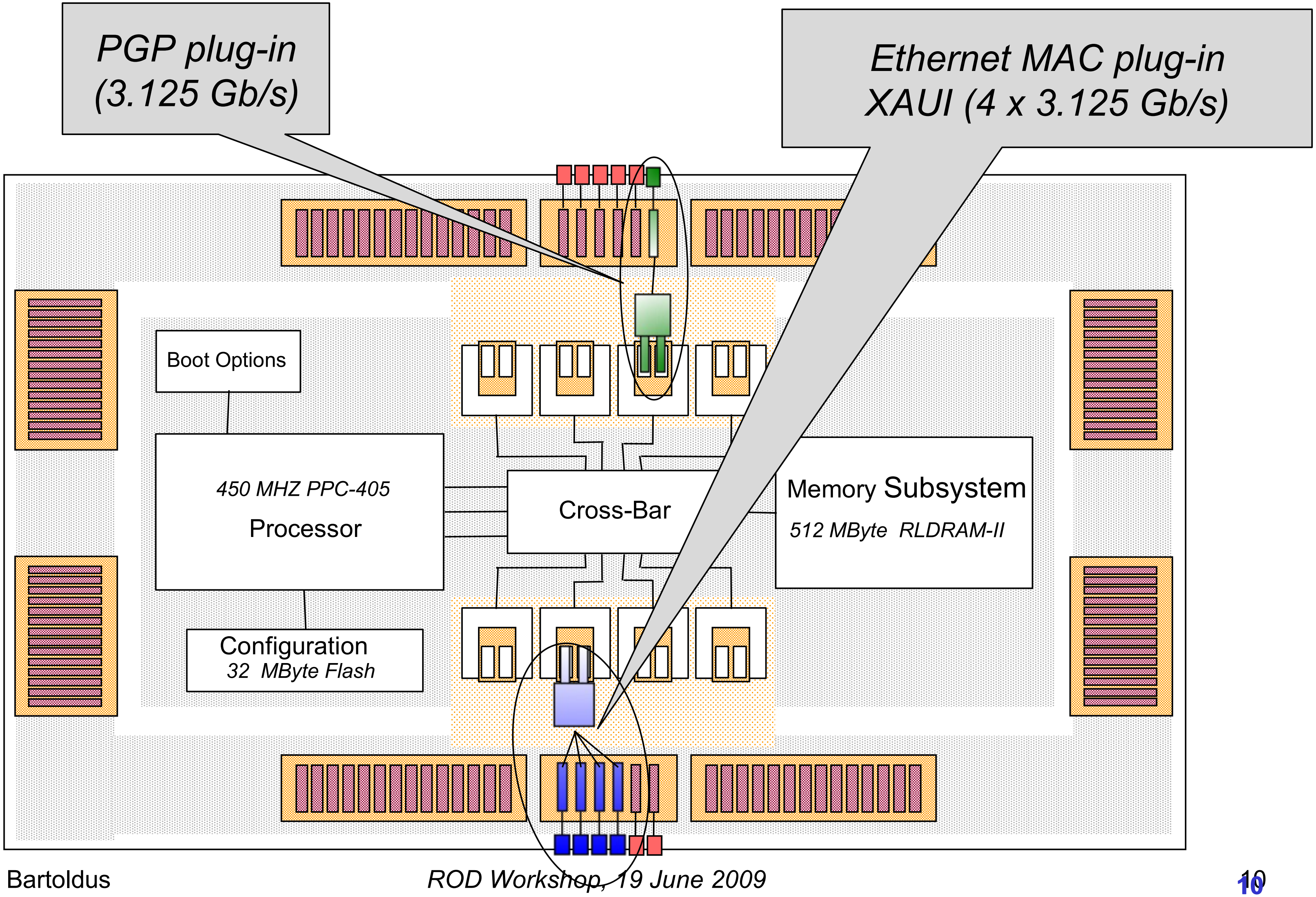
# Software & Development

- **Cross-Development...**
  - GNU cross-development environment (C & C++)
  - Remote (network) GDB debugger
  - Network console
- **Operating System Support...**
  - Bootstrap loader
  - Open-Source Real-Time Kernel (RTEMS)
    - POSIX-compliant interfaces
    - Standard IP network stack
  - Exception handling support
- **Object-Oriented Emphasis**
  - Class libraries (C++)
    - DEI support
    - Configuration interface

# Resources

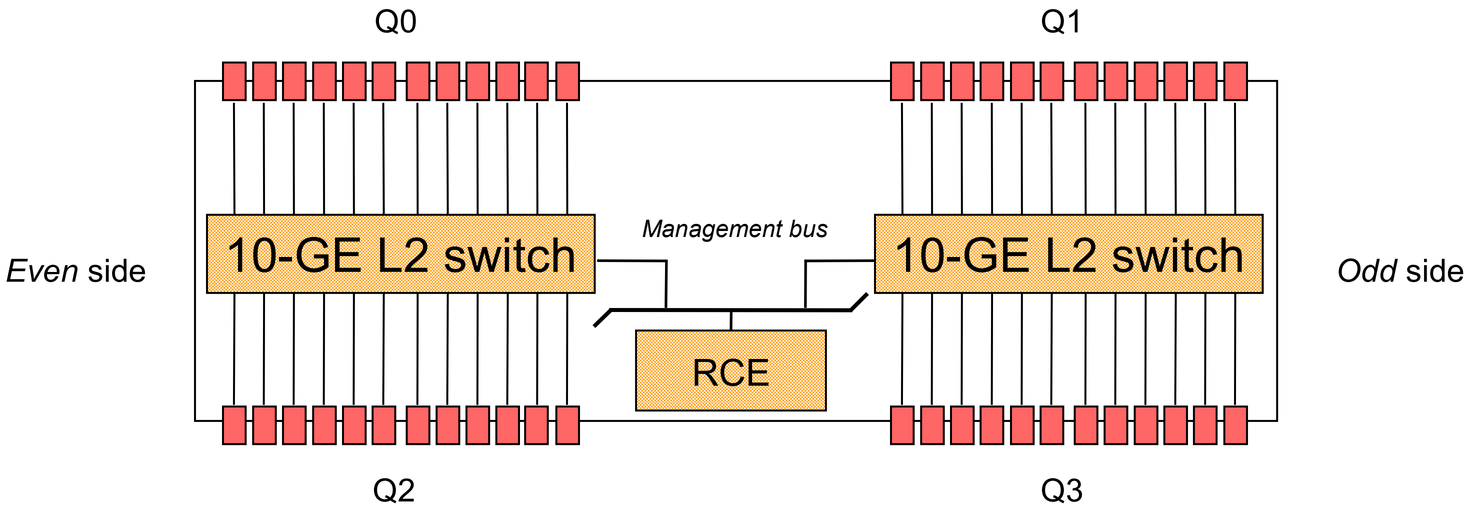
- **Multi-Gigabit Transceivers (MGTs)**
  - Up to 12 channels of
    - SER/DES
    - Input/output buffering
    - Clock recovery
    - 8b/10b encoder/decoder
    - 64b/66b encoder/decoder
  - Each channel can operate up to 6.5 Gb/s
  - Channels may be bound together for greater aggregate speed
- **Combinatoric Logic**
    - Gates
    - Flip-flops (block RAM)
    - I/O pins
- **DSP Support**
  - Contains up to 192 Multiple-Accumulate-Add (MAC) units

# “Plug-ins”



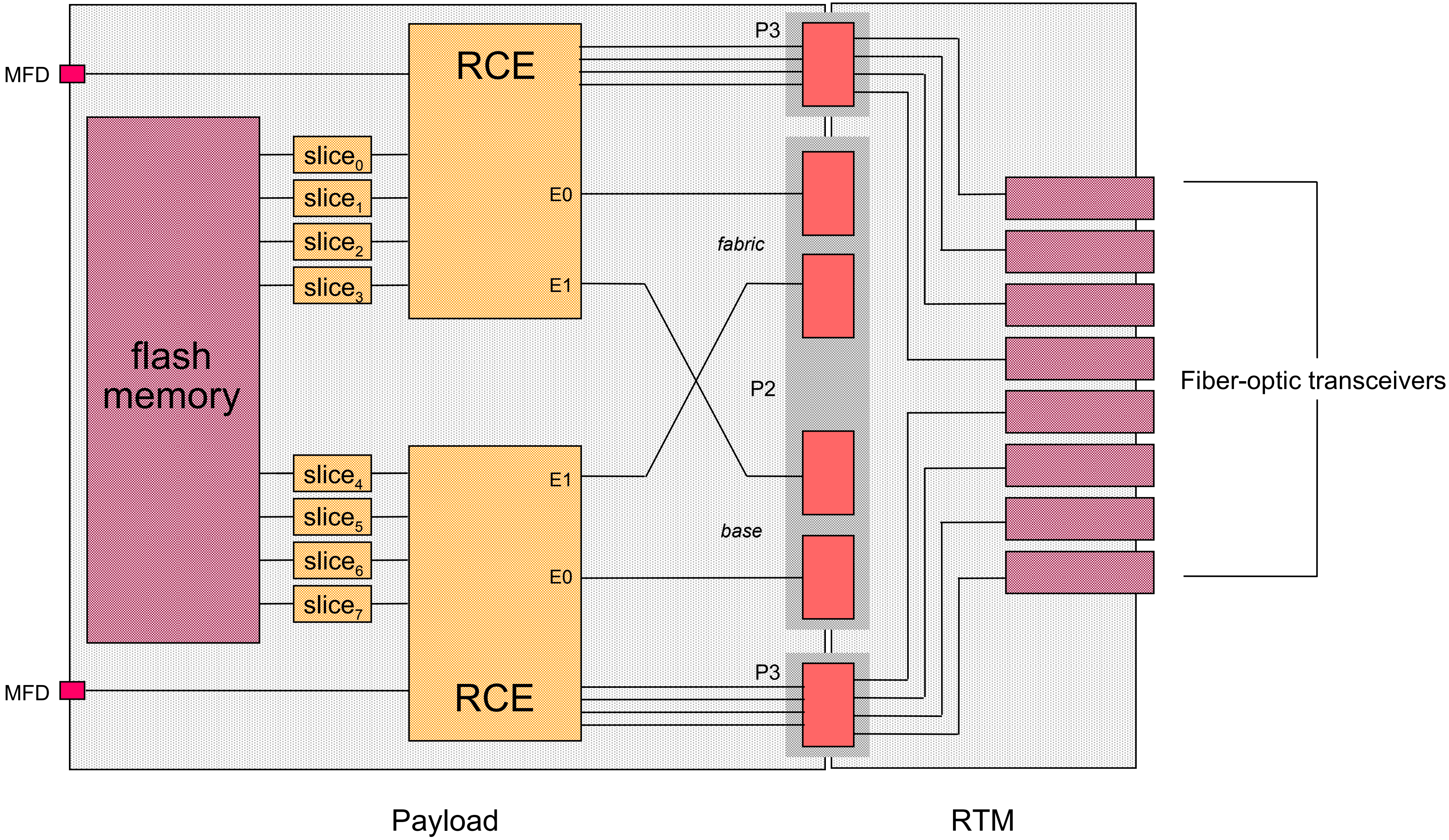
*PGP plug-in*
*(3.125 Gb/s)*
*Ethernet MAC plug-in*
*XAUI (4 x 3.125 Gb/s)*
Boot Options
*450 MHZ PPC-405*
Processor
Cross-Bar
Memory Subsystem
*512 MByte RLDRAM-II*
Configuration
*32 MByte Flash*

# The Cluster Interconnect (CI)

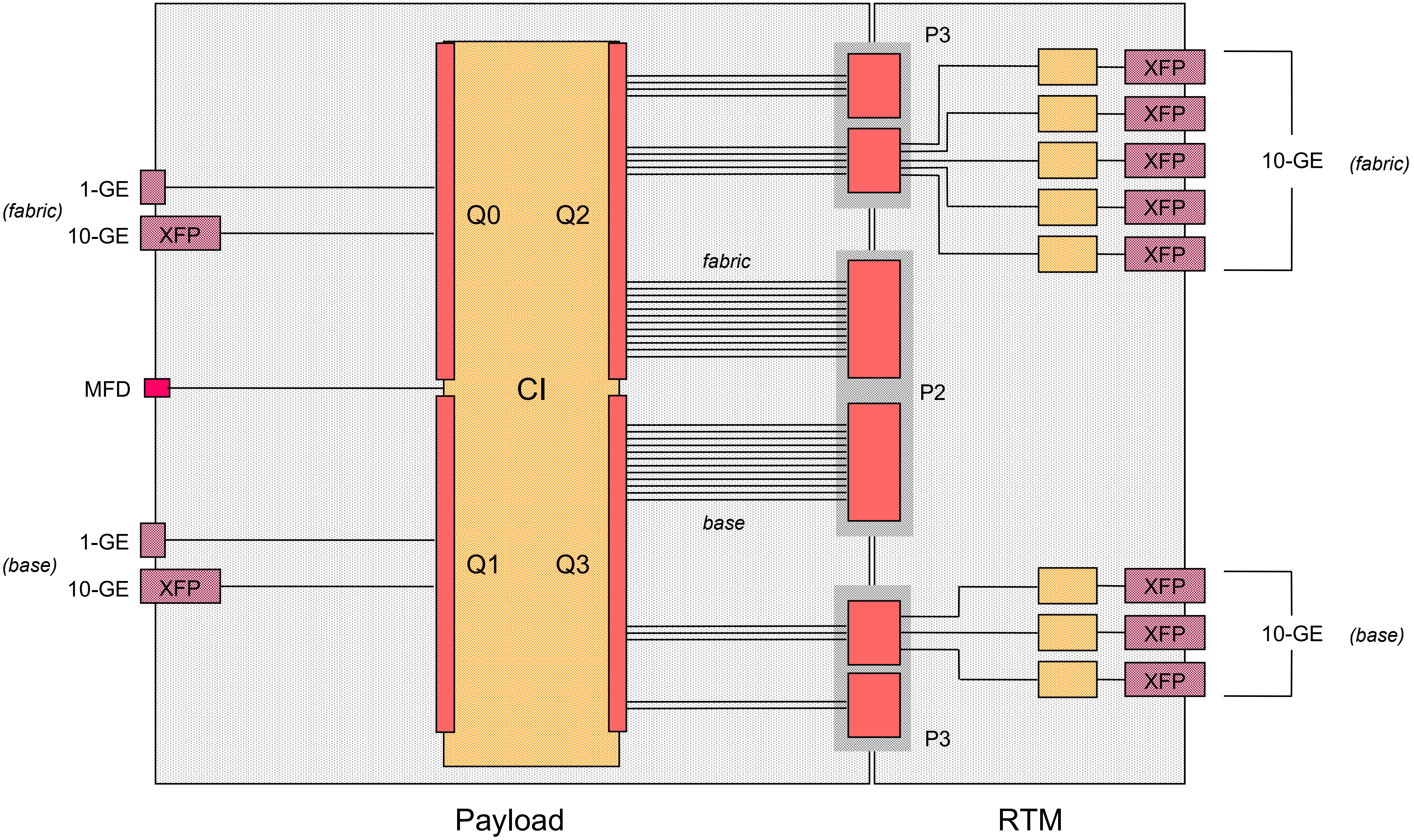


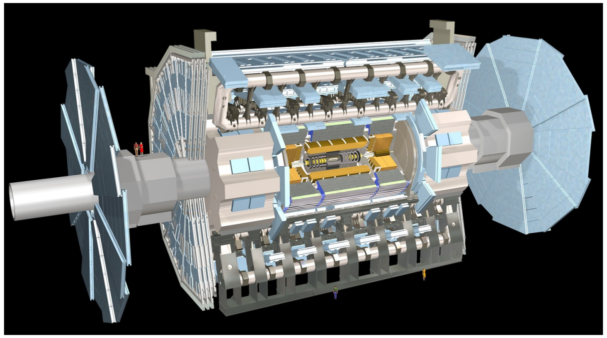
- **Based on two *Fulcrum* FM224s**
  - 24-port 10 GE switch
  - Is an ASIC (packaging in 1433-ball BGA)
  - 10-GE XAUI interface, however, supports multiple speeds
    - 100-BaseT, 1-GE, and 2.5 Gb/s
  - Less than 24 Watts at full capacity
  - Cut-through architecture (packet ingress/egress < 200 ns)
  - Full Layer-2 functionality (VLAN, multiple spanning tree etc.)
  - Configuration can be managed or unmanaged

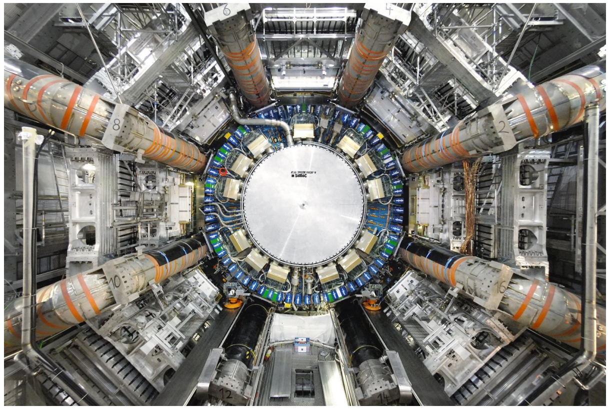
# RCE Board + RTM (Block Diagram)



# Cluster Interconnect Board + RTM

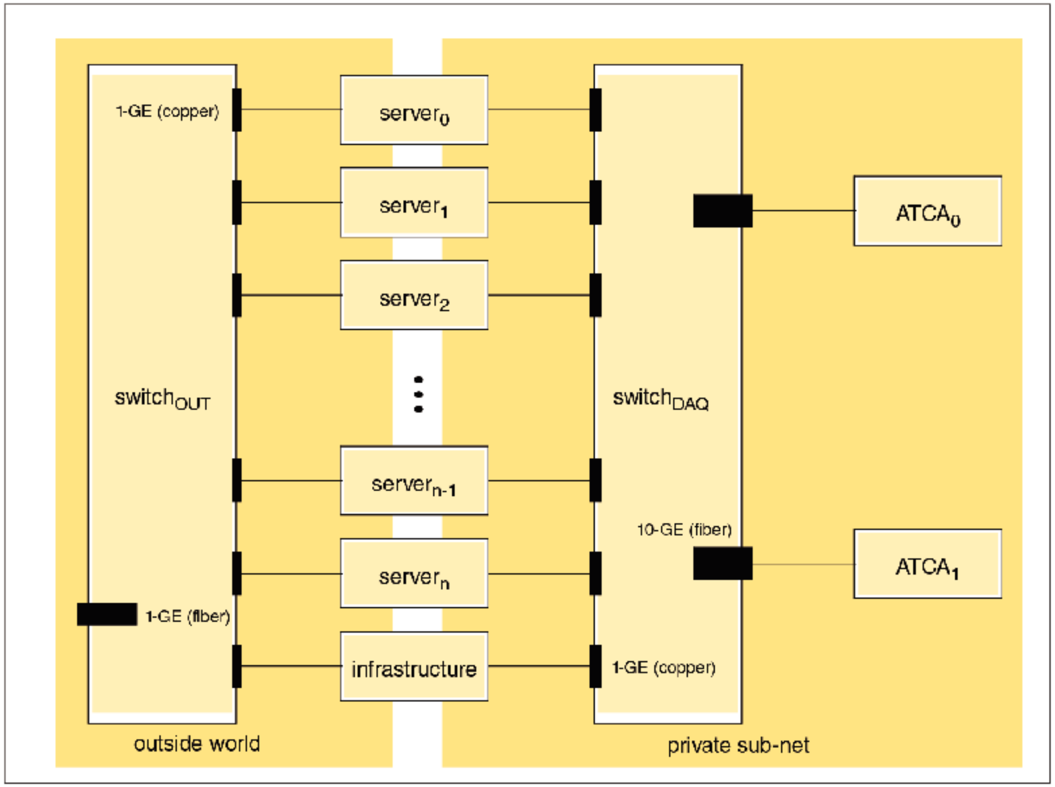






# RCE Development Lab

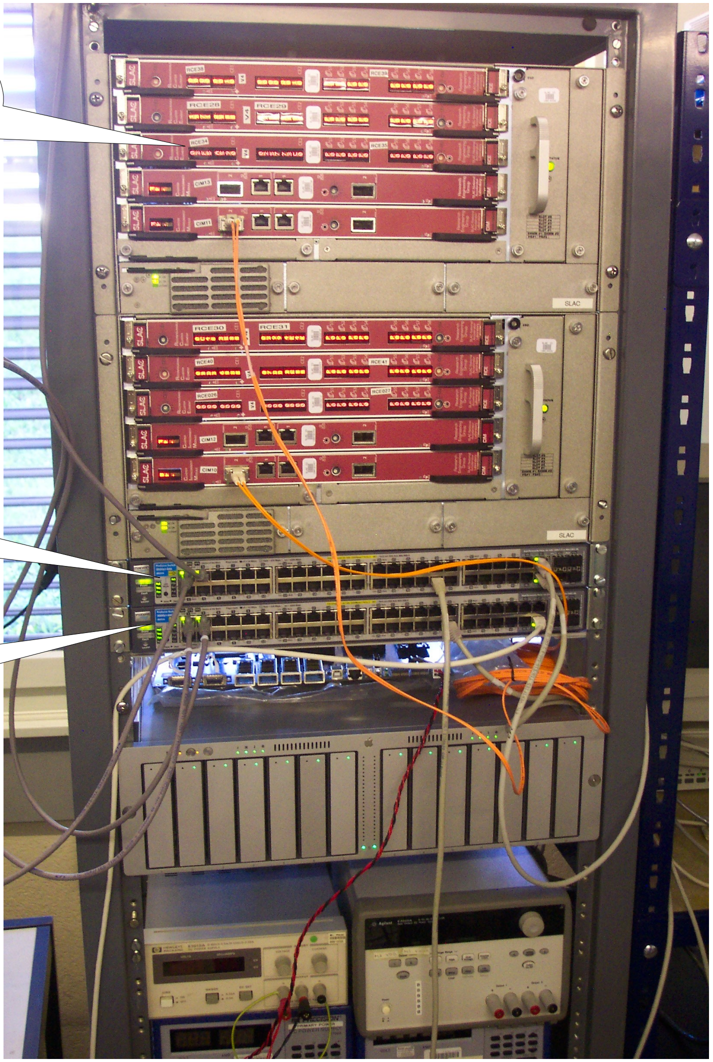
# Networking Setup



- **Two isolated Networks**
  - One public/one private
  - Infrastructure server and all development servers are dual-homed
  - ATCA crates only visible on private subnet

# RCE/CIM Development Platform

## Installation in Bldg 32



RCE/CIM
ATCA Crates

DAQ Switch
48x1GE + 2x10GE

GPN Switch
48x1GE

*HP ProCurve 3500yl*
*2x 10GE X2 SR*

Development/Test
Machines
2x 1GE NIC

*2x Dual-3GHz Xeon*

Infrastructure
Server for
DHCP, DDNS
NFS, NTP

*PowerEdge 2900*

# Acknowledgments

- **Many thanks to those at CERN who helped us put the teststand together in time for the RCE (and ROD) workshop**
  - Fred Wickens, (always a great help!) for pointing us in the right direction and for always knowing what we want
  - David Francis, for listening and finding the right people to talk to, and for moving the ROD workshop
  - Stefan Stancu, for buying our switches and helping to install them in no time, and for many little things
  - Marc Dobson, for helping to set up the network and the PCs and for much good advice
  - Gokhan Unel (*), for finding and lending us the two PCs
  - Haimo Zobernig, Werner Wiedenmann, Neng Xu, Andre dos Anjos, and Sau Lan Wu, for kindly letting us use their rack in Lab32 and for helping with power and sudden a/c failures

(*) PLEASE BE PATIENT!

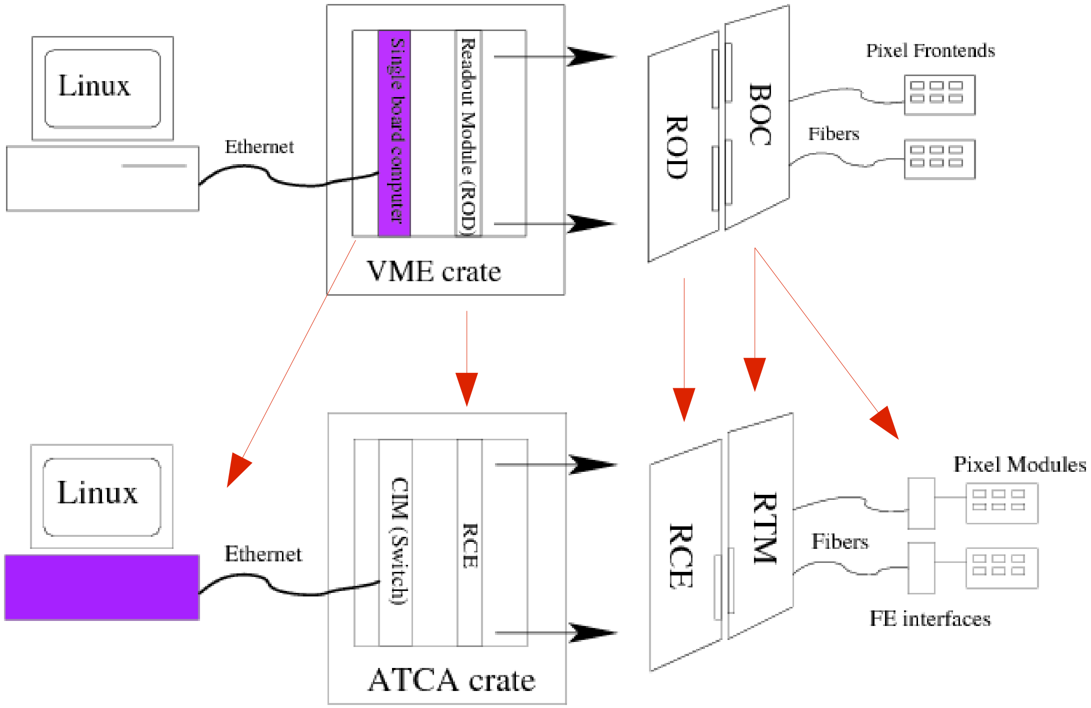
# RCE Training Workshop

- **Took place Monday and Tuesday of this week**
  - 33 participants
- **Tutorials, Demonstrations**
- **Discussion Towards Future Collaborations**
- **Hands-On Session**
  - 18 people requested accounts (and counting)
  - Everyone learned to develop and execute “Hello World!”
  - Some learned quite a bit more than that...
- **After the Workshop**
  - Accounts are still in use
  - 12 RCEs can be used independently
    - Contact us if you are interested in exploring one
  - This training focused specifically on software
  - Plan to have a future workshop on FPGA firmware

# Case Study: Pixel FE Calibration

*Current ATLAS Setup*

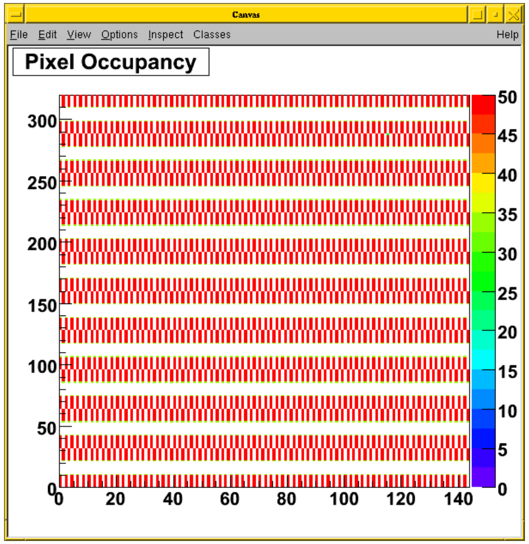
*Martin Kocian's Setup*



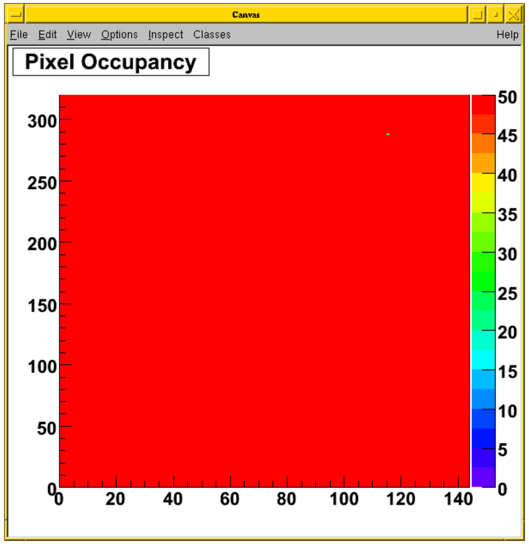
- **Pixel FE Digital Test was ported from Pixel ROD to the RCE**
  - Modified DSP code runs on PowerPC processor
  - Controlled by a linux host that communicates with the RCE
  - Front-End communication through fiber at 3.125 Gb/s
  - Runs successfully...

# Pixel FE Digital Test on the RCE

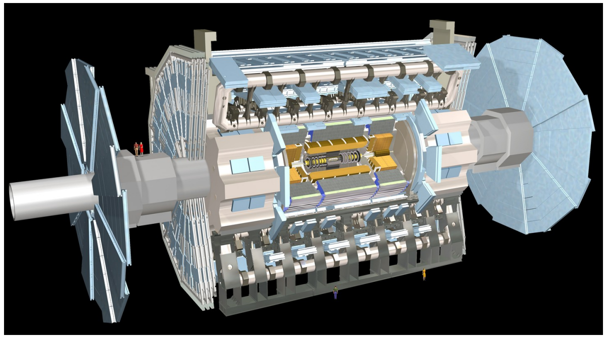
- **Pixel Digital Test runs on the RCE**
  - Martin ran his setup in front of the workshop audience on Tuesday
  - This is a concrete example of replacing the BOC/ROD/SBC(VME) chain with RTM/RCE/CIM(ATCA) and a Linux host
  - The original DSP code could be ported without major changes
    - Only complication was byte-swapping between PowerPC (big-endian) and DSP (little-endian)
- **This example focused on reproducing functionality**
- **Future test cases to explore & compare bandwidth**



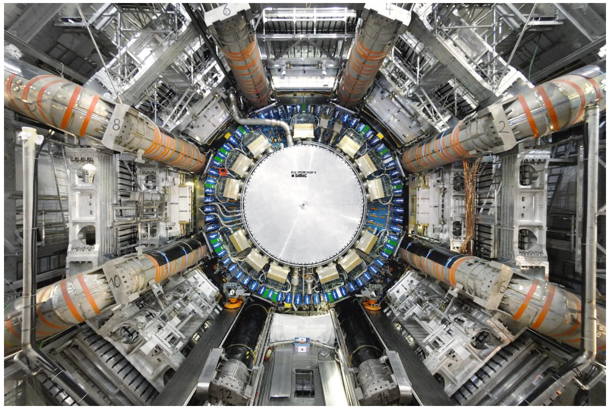
running...



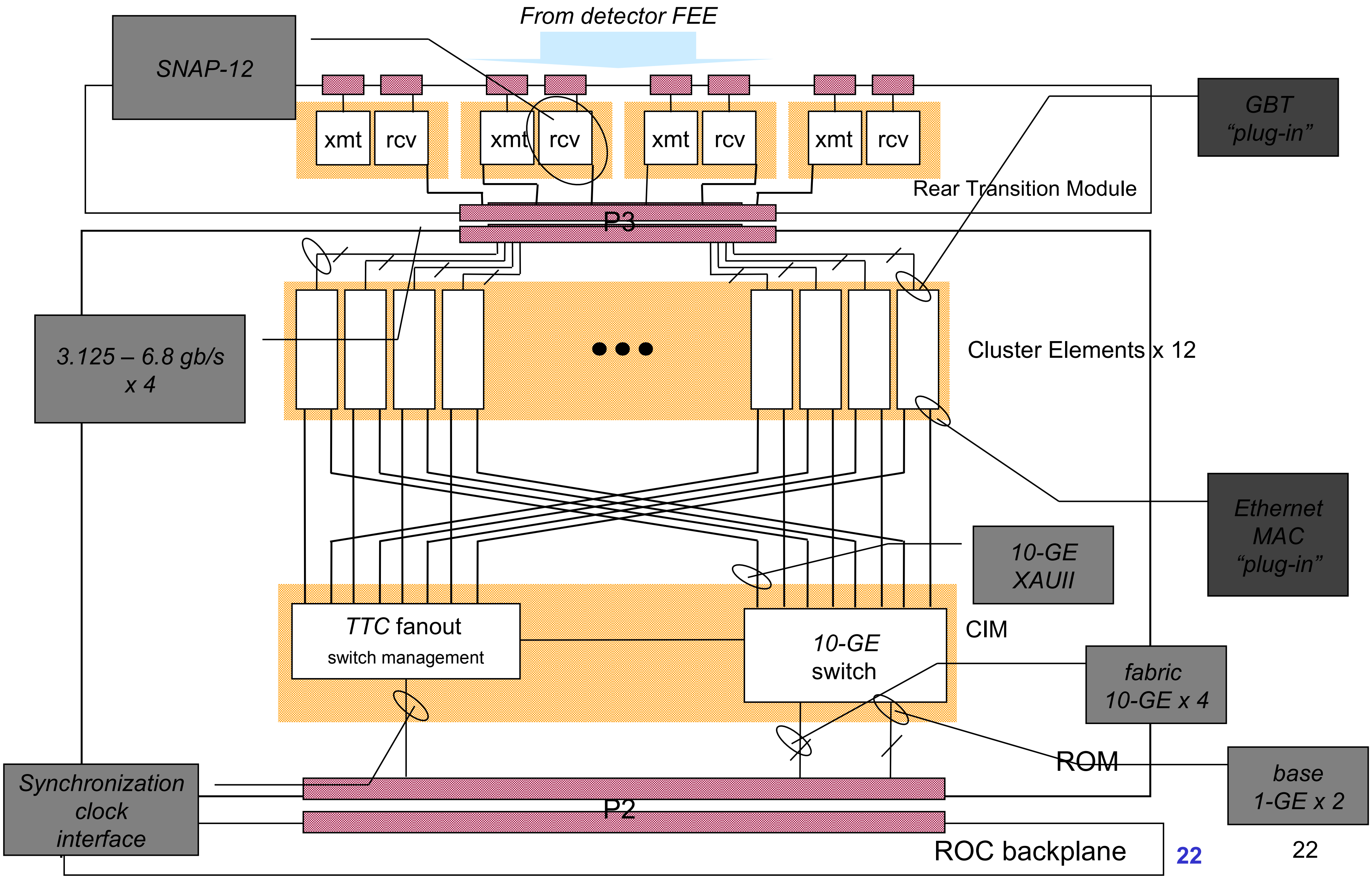
finished



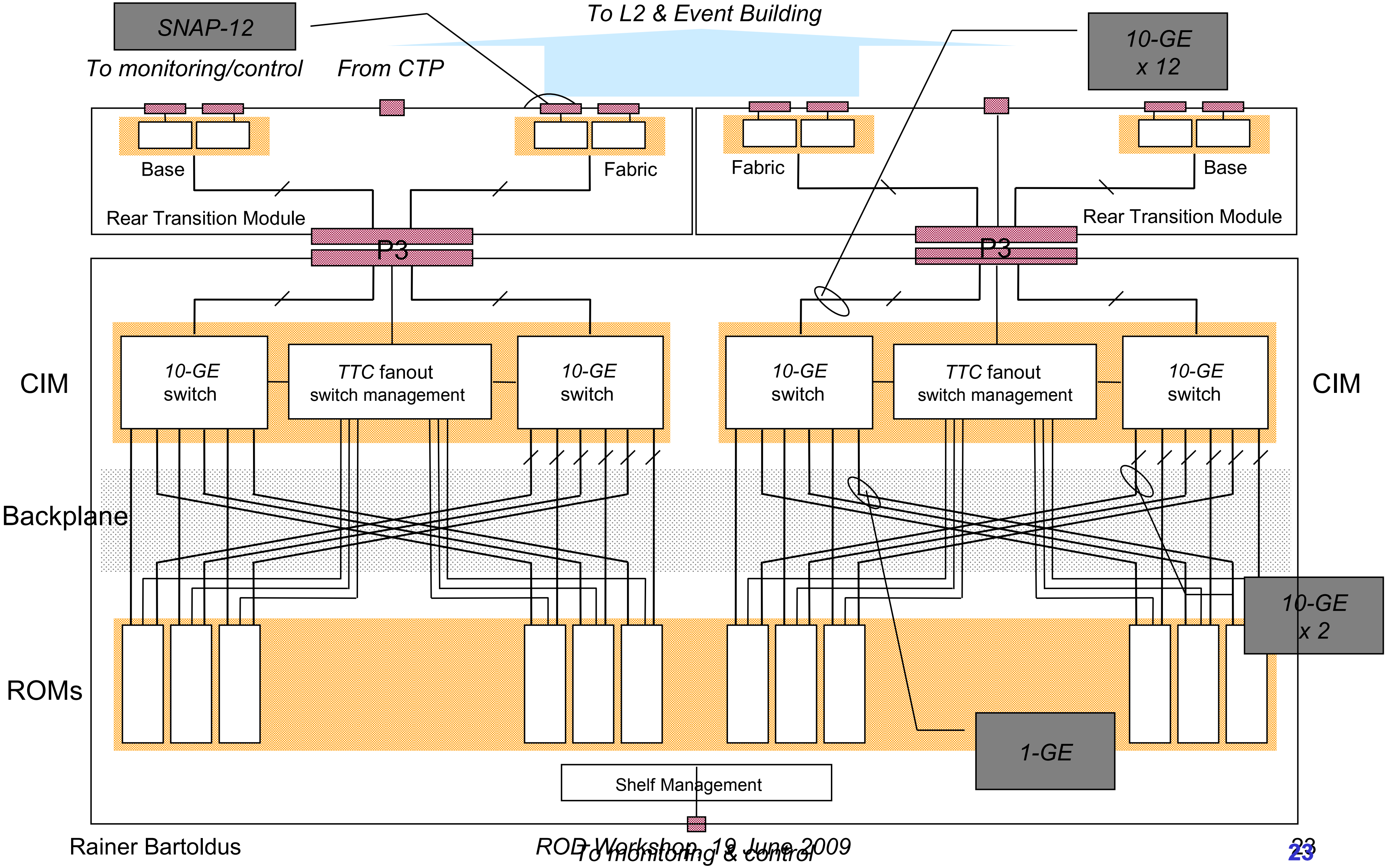
# A Hypothetical 48-channel Read-Out Module



# Hypothetical 48-Channel Read-Out Module



# Hypothetical Read-Out Crate



# Current (Future) RCE Bandwidth, Toy Model

- **Future ROM Board**
  - Assumed to be able to host 12 RCEs in 6 FPGAs
- **RTM Back Panel**
  - Going from *XFP* to *Snap-12* should fit 2x8x12 fibers, in pairs of up and down (full duplex), or 96 detector channels
  - GBT at 5 GHz delivers 3.2 Gb/s usable bandwidth (may hope for 6.4 Gb/s at 10 GHz ?)
    - 307 Gb/s (614 Gb/s) per RTM
- **ATCA Zone 3 Connector**
  - Can go up to 400x2 differential pairs and run 3 Gb/s per pair today (possibly 10 Gb/s in the future) so not a bottleneck
- **RCE/ROM Input**
  - Four MGT lanes per RCE, 6.8 Gb/s today (10 Gb/s in the future)
    - 27 Gb/s (40 Gb/s) per RCE
    - 326 Gb/s (480 Gb/s) per ROM
      - Matches today's GBT rates (would be limiting factor in the future)

# Current (Future) RCE Output

- **RCE Processing Power**
  - Implementation dependent, and on how much can be offloaded to DSPs or gates
    - 30-40 Gb/s per RCE or 360-480 Gb/s per ROM might be a good guess, which matches the input bandwidth
- **RCE Output**
  - Four lanes of MGT or 10 Gb/s (10 GE)
    - If FEX does 1:3 to 1:4 data reduction this can serve data at L1 rate
- **ROM Crate (ROC) Output**
  - ROM output is 2 x 10 Gb/s for each star
    - 40 Gb/s per ROM on dual star
      - (15 % of L1 for tracker with in:out = 1:1 and more for calorimeter with a few times data reduction)
  - Each star supports up to 6 boards
    - 240 Gb/s per ROC

# Future DAQ Plant on a Napkin

- **Balance of Real Estate and Throughput**
  - For silicon tracker may not push GBT beyond 3.2 Gb/s and RTM to full density but rather add more ROMs to scale to more than 15% L1 accepts for L2
  - For calorimeter with data reduction of a few, keep minimum plant size with still a large fraction of L1
  - Interesting case: with data reduction of 1:10, one can serve full L1 rate to L2
- **LAr ROMs**
  - Input 1500 FEBs of 100 Gb/s each; at 6.4 Gb/s GBT this is 16 fibers per FEB or 6 FEBs per ROM or 250 ROMs
- **Pixel ROMs**
  - SLHC Pixel has 18x data rate of current detector or 800 fibers of 3.2 Gb/s
    - With 48 fibers per ROM (half) one has an 18 ROM system that could output 20% of the L1 accepts
    - Compare that to 134 ROD plus 12 ROS system

# Summary (1/2)

- **We have started to explore a new generation DAQ platform**
  - Strategy is based on the idea of modular building blocks
    - Inexpensive computational element (the RCE)
    - Interconnect mechanism (the CI)
    - Industry standard packaging (ATCA)
  - Architecture is now relatively mature
    - Both demo boards (and corresponding RTMs) are functional
    - RTEMs ported and operating
    - Network stack fully tested and functional
  - Performance and scaling meet expectations
  - Documentation is a "work-in-progress"
- **This technology strongly leverages off industry innovation**
  - System-On-Chip
  - High speed serial transmission
  - Low-cost, small footprint, high-speed switching (10 GE)
  - Packaging standardization (serial backplanes and RTM)

# Summary (2/2)

- **Gained experience with these innovations will itself be valuable**
- **This technology offers a ready-today vehicle to explore both alternate architectures and different performance regimes**
  - A test platform has been installed with RCE/CIM building blocks mounted on prototype ATCA boards
  - Workshop material is available that gives a tutorial of the platform and its use for developers
  - As an example, the Pixel FE test was ported to and run on the RCE
  - “Back of the napkin” calculations indicate that this platform is on a possible trajectory towards a phase-II ROD/ROS
- **A teststand is available at CERN and everyone is welcome to explore!**