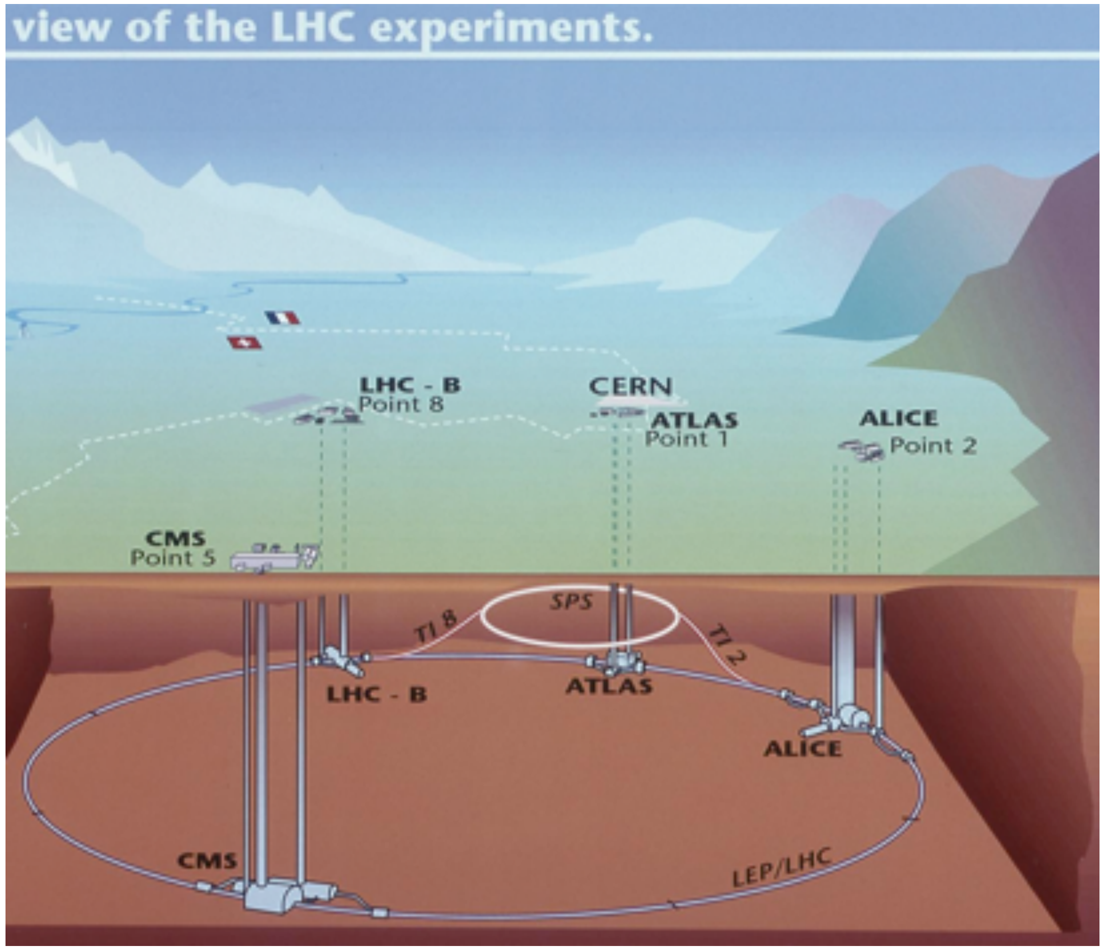
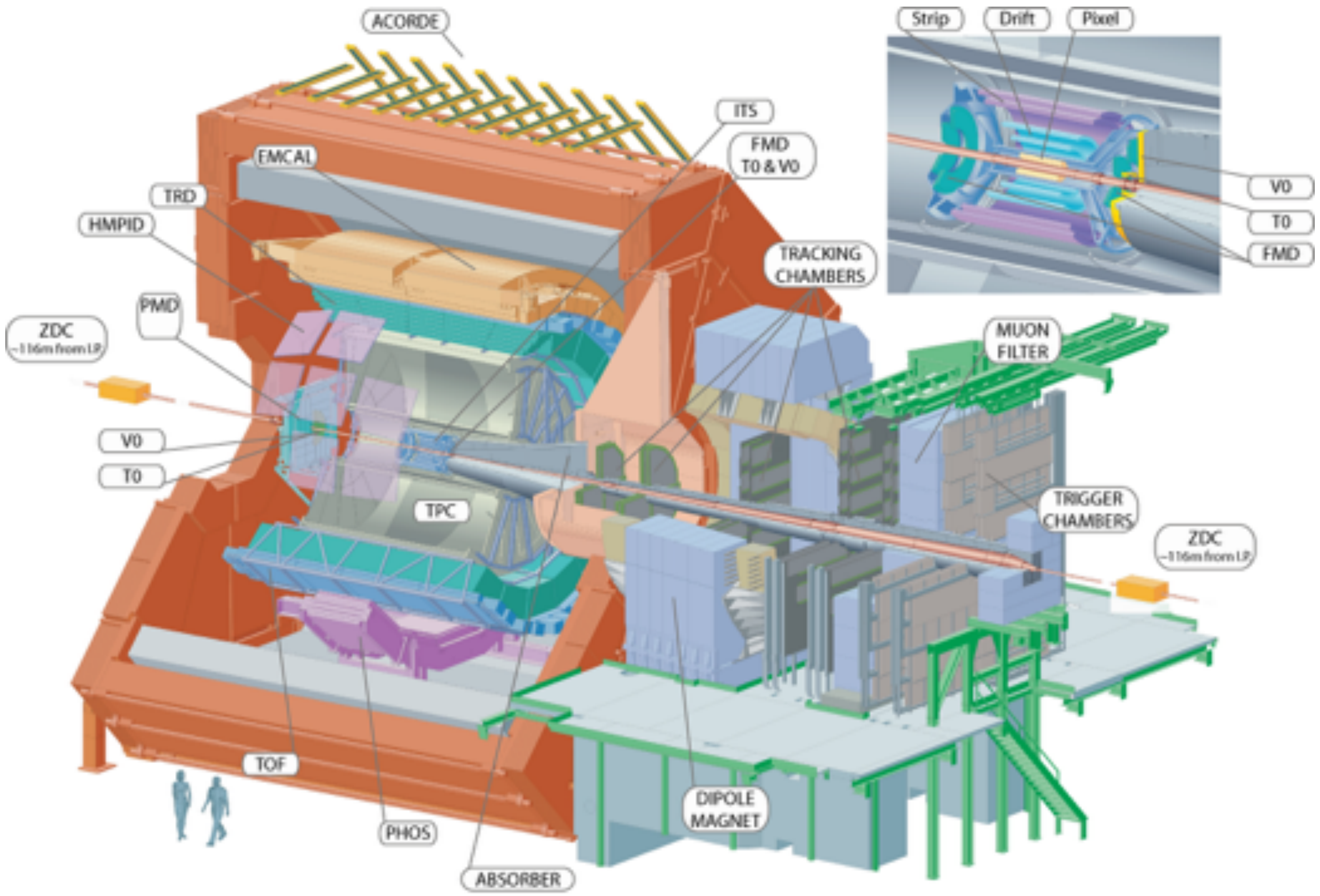
# ALICE High Level Trigger

## status and plans

M.Krzewicki for the ALICE collaboration

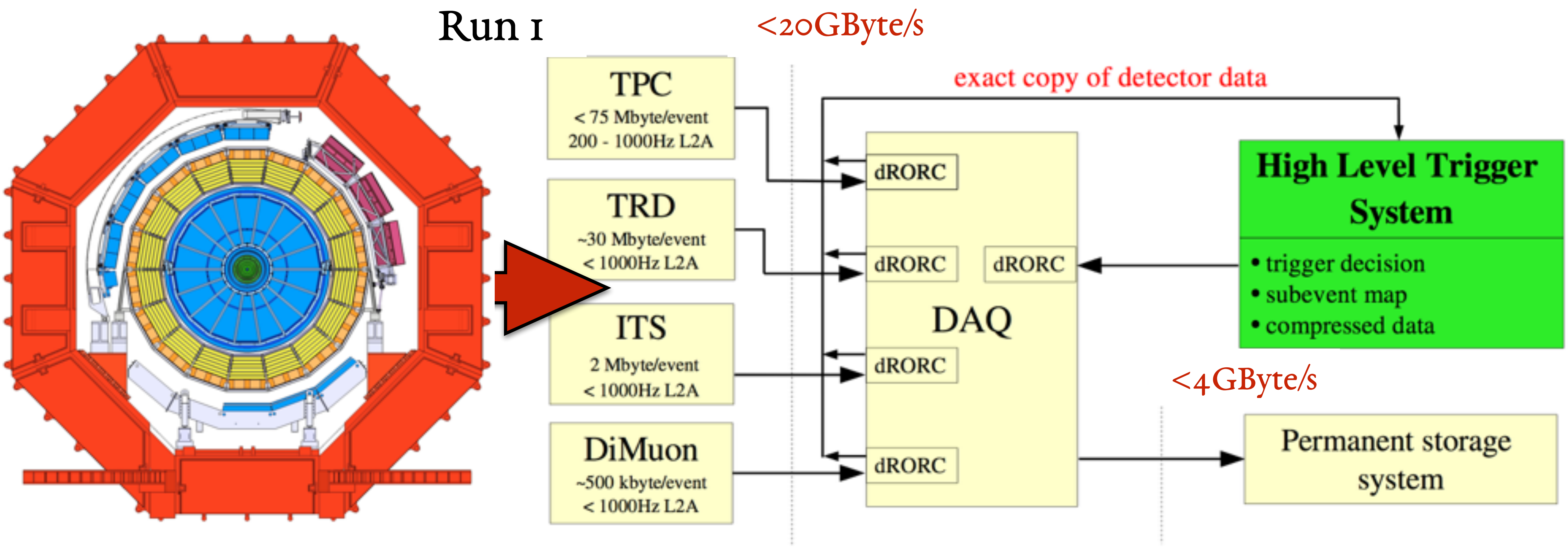
# The ALICE experiment





- A CERN experiment @LHC
- Optimised for heavy-ion data, takes also proton-proton.
- Located at LHC Point2 (St.Genis).

# The data flow



- Run2:
  - Higher interaction rates (8-30kHz PbPb).
  - Readout upgrade to RCU2 + DDL2: ~2x more bandwidth.
  - HLT evolved to comply with new constraints.

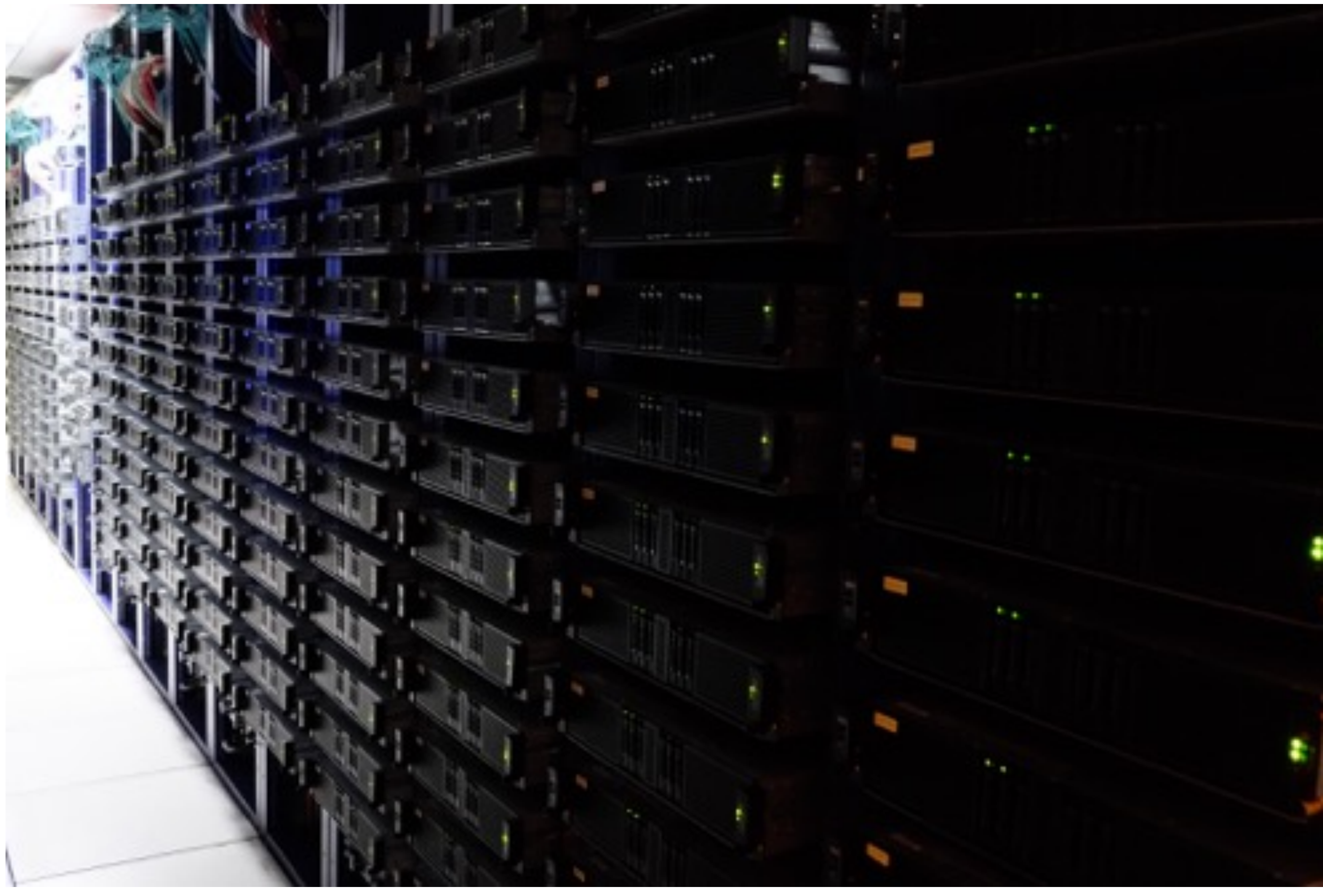
# The ALICE High Level Trigger

- Part of the old production system -> development cluster (~60 nodes).
  - Software development and validation.

- The new farm:
  - New layout (1 row of racks instead of 3).
    - Smaller is better - cheaper, less cable, easier to cool, ...
    - Homogeneous system - each node capable of all roles (FEP,CN).
    - Grouped in building blocks a 36 machines (3 racks).
    - All machines on UPS.
- Heavy utilisation of hardware acceleration: FPGA + GPU.
- Primary functions:
  - Data compression.
  - Online event reconstruction.
  - Online calibration.

# The ALICE High Level Trigger

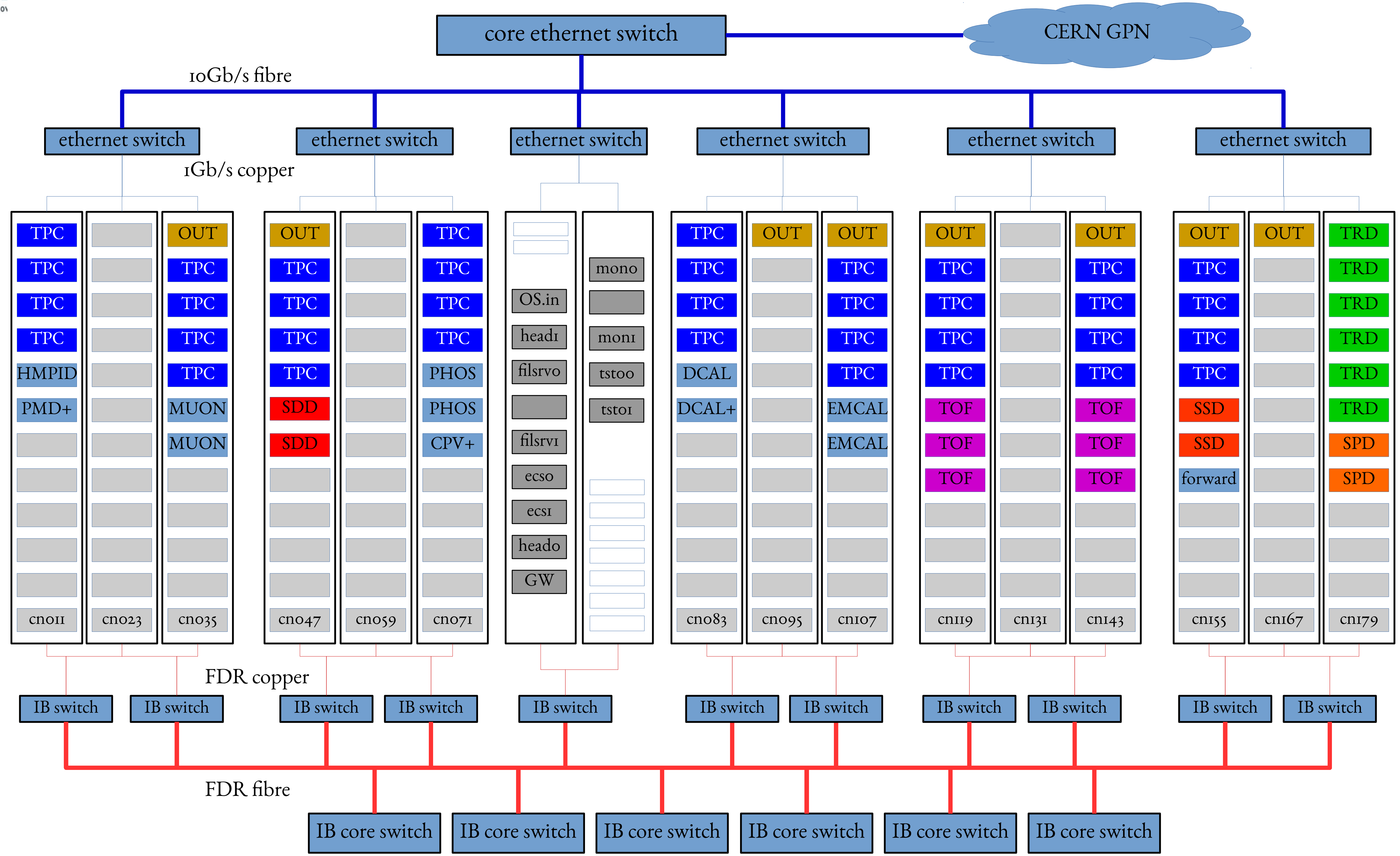
- 180 nodes - 4320 CPU cores:
  - 2x Intel Xeon E5-2697 CPUs (2.7 GHz, 12 Cores each).
  - 128 GB RAM.
  - 2x 240 GB SSD (used in Raid 1 - Mirroring).
  - 1 AMD FirePro S9000 GPU.
  - 1 C-RORC board (installed in 74 nodes).



- 6+ Infrastructure Nodes:
  - 2x Intel Xeon E5-2690, 3.0 GHz 10 Cores.
  - 128 GB RAM.
  - 2x 240 GB SSD (Raid 1 - mirroring).

- Network:
  - Data: Infiniband in IPoIB Mode ( FDR with 56Gb/s, full bisection bandwidth).
  - Management: gigabit ethernet with sideband IPMI - one physical ethernet port per node.
    - 10Gbit backbone.

# The anatomy of the ALICE HLT



core ethernet switch

CERN GPN

10Gb/s fibre

ethernet switch

ethernet switch

ethernet switch

ethernet switch

ethernet switch

ethernet switch

1Gb/s copper

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cn011 | TPC | TPC | TPC | TPC | HMPID | PMD+ | | | | | |
| cn023 | | | | | | | | | | | |
| cn035 | OUT | TPC | TPC | TPC | TPC | MUON | MUON | | | | |
| cn047 | OUT | TPC | TPC | TPC | TPC | SDD | SDD | | | | |
| cn059 | | | | | | | | | | | |
| cn071 | TPC | TPC | TPC | TPC | PHOS | PHOS | CPV+ | | | | |
| | | | OS.in | head1 | filsrv0 | | filsrv1 | ecs0 | ecs1 | head0 | GW |
| | | mono | | mon1 | tst00 | tst01 | | | | | |
| cn083 | TPC | TPC | TPC | TPC | DCAL | DCAL+ | | | | | |
| cn095 | OUT | | | | | | | | | | |
| cn107 | OUT | TPC | TPC | TPC | TPC | EMCAL | EMCAL | | | | |
| cn119 | OUT | TPC | TPC | TPC | TPC | TOF | TOF | TOF | | | |
| cn131 | | | | | | | | | | | |
| cn143 | OUT | TPC | TPC | TPC | TPC | TOF | TOF | TOF | | | |
| cn155 | OUT | TPC | TPC | TPC | TPC | SSD | SSD | forward | | | |
| cn167 | OUT | | | | | | | | | | |
| cn179 | TRD | TRD | TRD | TRD | TRD | TRD | SPD | SPD | | | |

FDR copper

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

IB switch

FDR fibre

IB core switch

IB core switch

IB core switch

IB core switch

IB core switch

IB core switch

- building block: 3 racks +{eth.,IB,IB} switch.
- Coloured blocks: 74 machines equipped with detector readout/DAQ interface (C-RORC).

# Provisioning and configuration

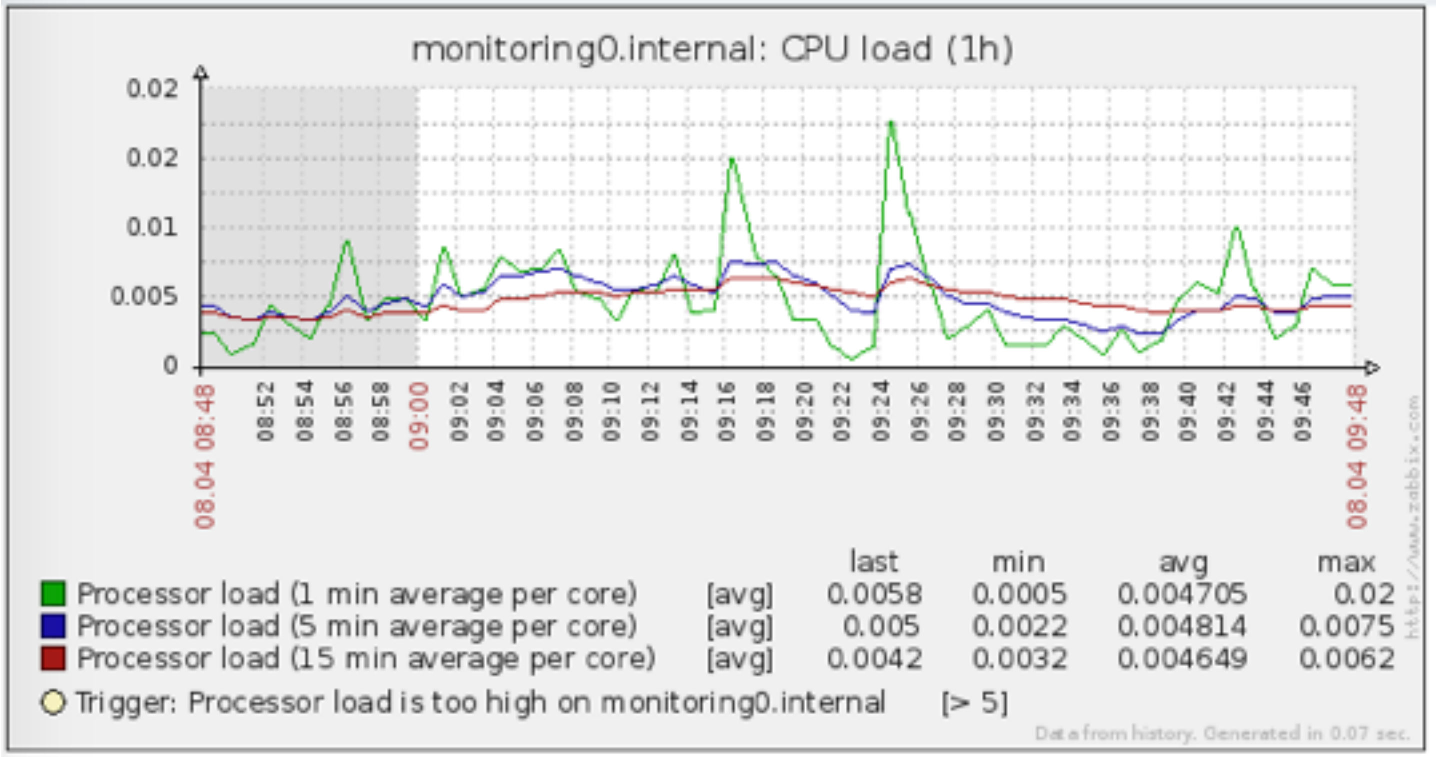
- Foreman:
  - DHCP,DNS,TFTP.
  - Puppet.
- Nodes: network boot (PXE).
  - Local/ethernet boot.
- Puppet:
  - all of local node configuration

- Current OS: Fedora 20
  - contemplating others (CentOS7)

- Full config/deployment automation
  - production system
  - development farm

# Monitoring

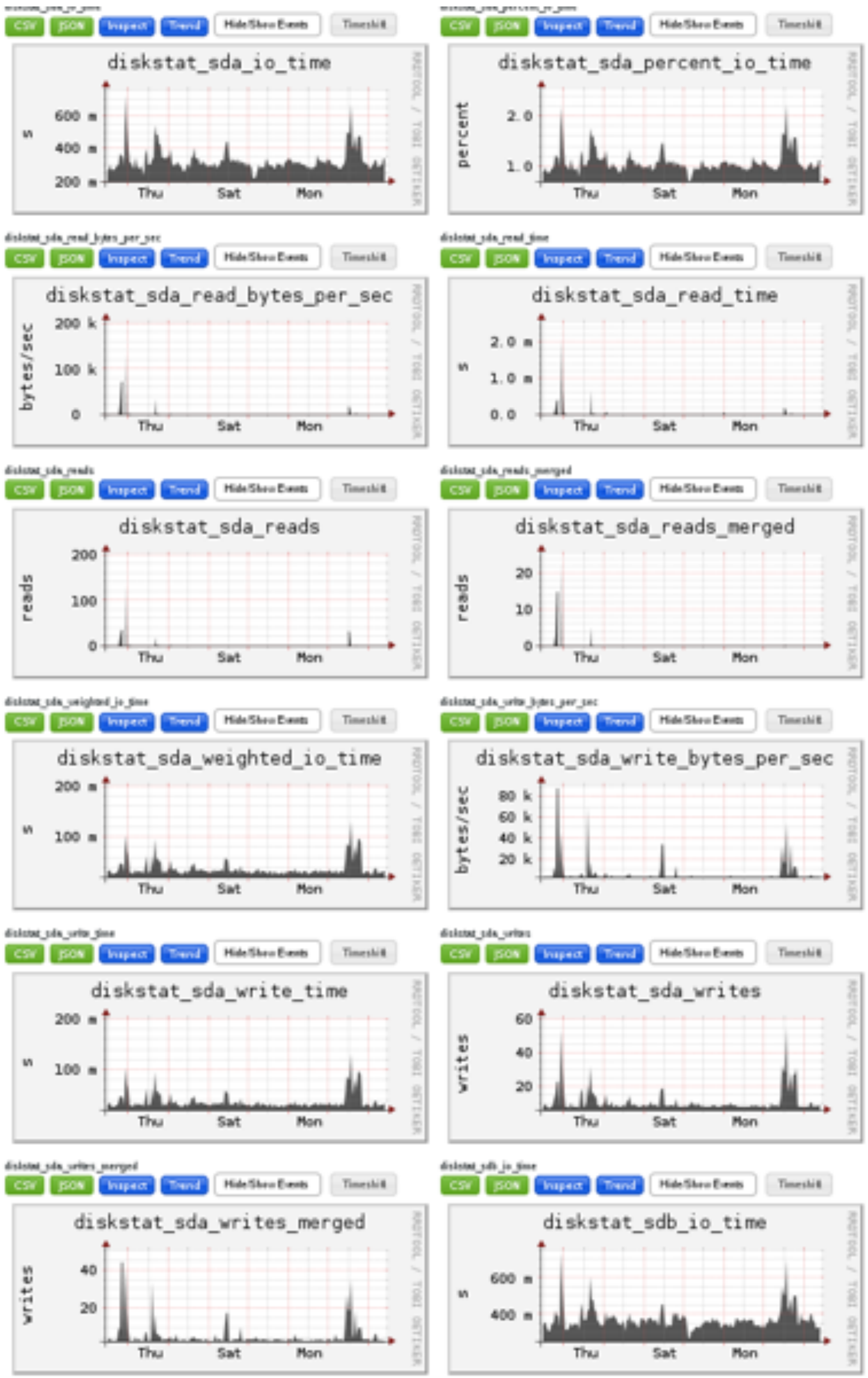
- Zabbix+Ganglia.
  - All compute nodes.
  - All servers.
  - Centralised, easy to use.
  - Automated alarms, trends, ...
- Automatic deployment (puppet+foreman).



| | | last | min | avg | max |
|---|---|---|---|---|---|
| 🟩 Processor load (1 min average per core) | [avg] | 0.0058 | 0.0005 | 0.004705 | 0.02 |
| 🟦 Processor load (5 min average per core) | [avg] | 0.005 | 0.0022 | 0.004814 | 0.0075 |
| 🟥 Processor load (15 min average per core) | [avg] | 0.0042 | 0.0032 | 0.004649 | 0.0062 |

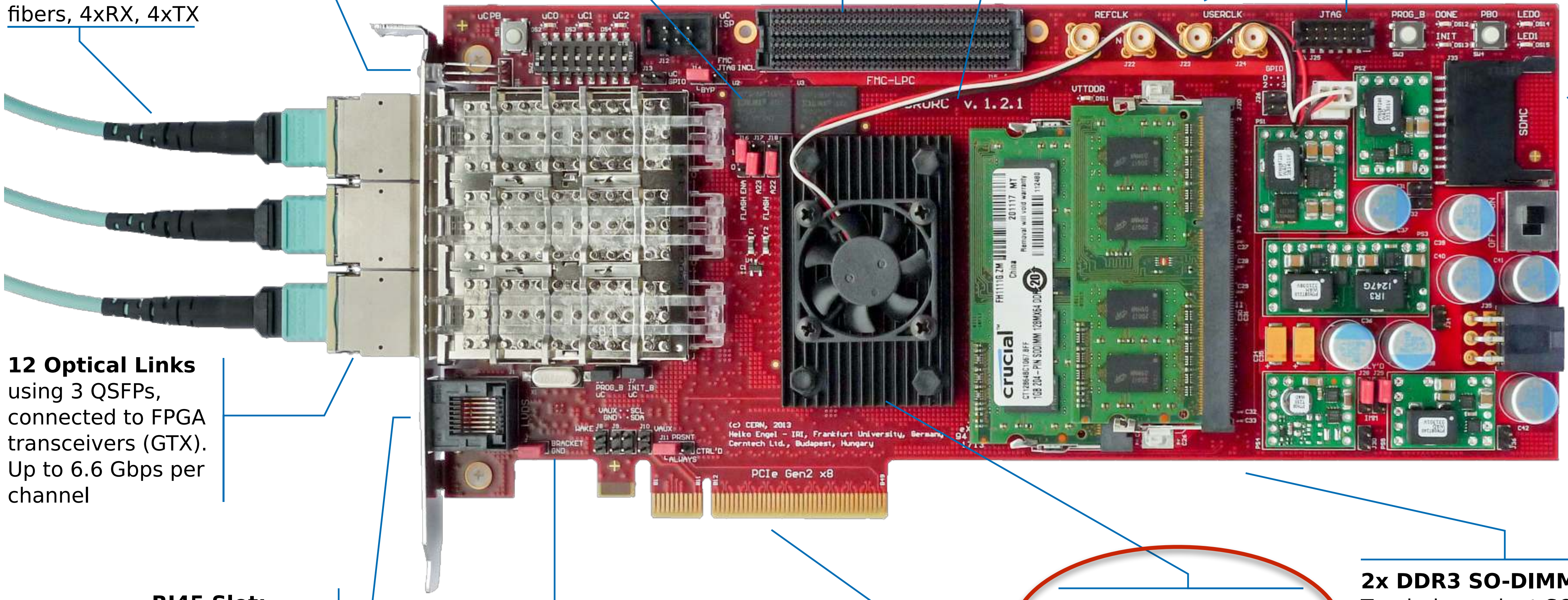
○ Trigger: Processor load is too high on monitoring0.internal [> 5]

Data from history. Generated in 0.07 sec.



# Common ReadOut Receiver Card

## C-RORC Overview



**MPO/MTP Fibers**
Parallel optical
fibers, 4xRX, 4xTX

Bi-Color **LED**
in IO-Bracket

**Configuration Flash**
2x 128Mbit Xilinx
PlatformFlash XL

**FMC LPC**
VITA57.1 FMC
LPC connector

**GTX RefClk**
Configurable GTX
reference clock

**System Clock**
200 MHz

**SMA Connectors**
for additional GTX
and system clocks

**JTAG-Connection**
for programming
cable

**SDCard Socket**

**Power Connection**
via 6-Pin PCIe GPU
power connector

**12 Optical Links**
using 3 QSFPs,
connected to FPGA
transceivers (GTX).
Up to 6.6 Gbps per
channel

**RJ45 Slot:**
4 LVDS pairs,
up to 800 Mbps.
Not for Ethernet

**Microcontroller** for
configuration monitoring
and connection to host via
PCIe SMBus lines

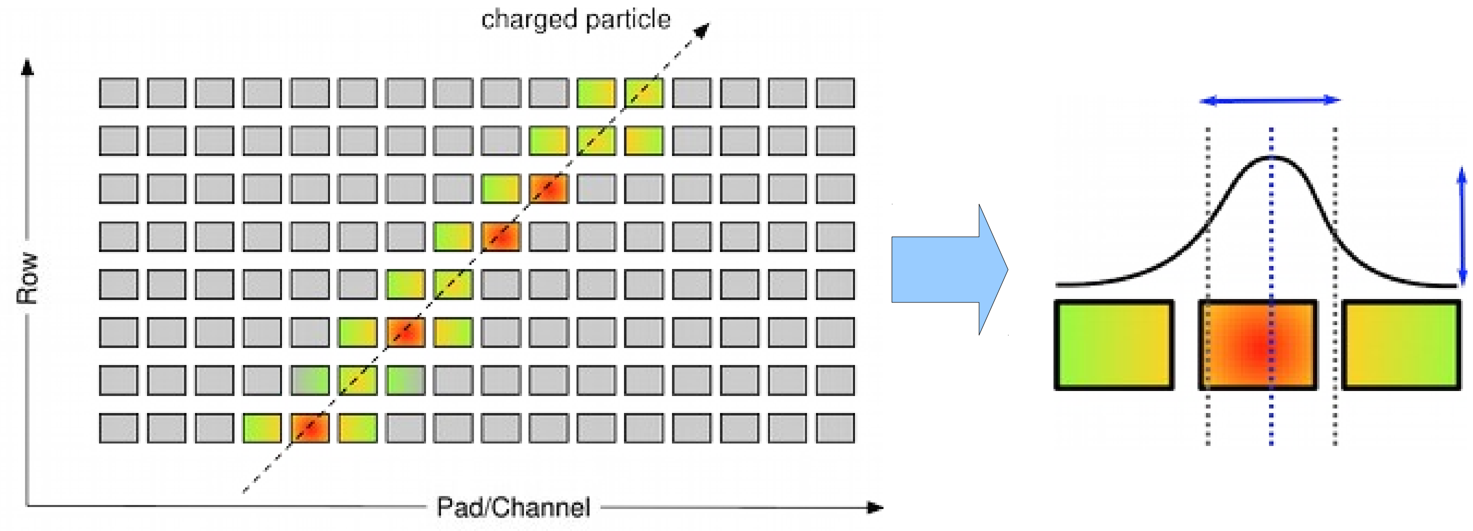
**PCIe Gen2, 8 Lanes**
8x 5.0 Gbps, connected to
Xilinx PCIe Hard Block

**Xilinx Virtex-6 FPGA**
FF1156 package
LX240T

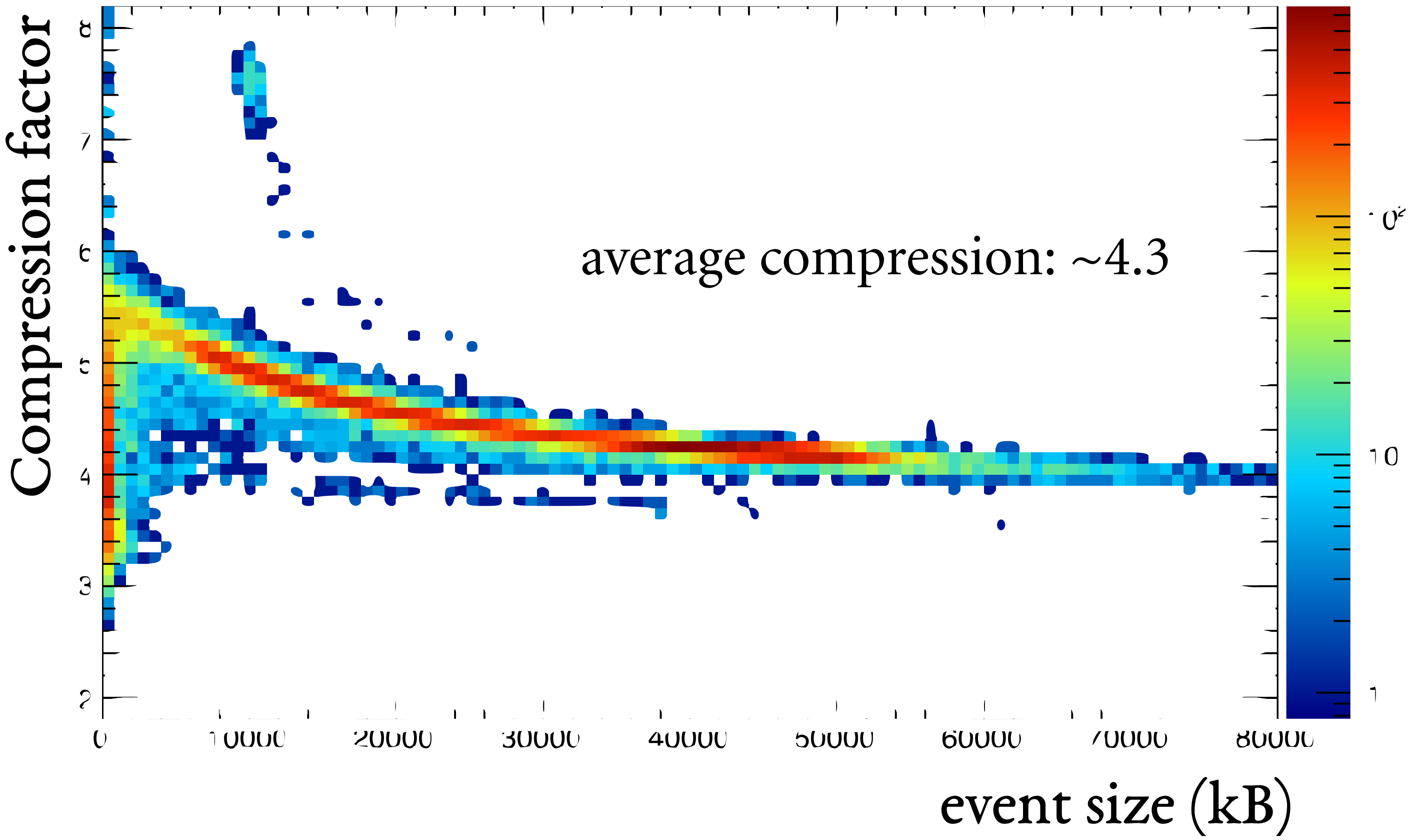
**2x DDR3 SO-DIMM**
Two independent SO-DIMMs
up to 1066 Mbps with Xilinx
MIG DDR3 softcore

- Used in ALICE DAQ+HLT *and* ATLAS TDAQ ROS.
- Up to 12 optical DDL links (DDL 1 & 2), 74 cards total, ~500 links.
- FPGA hosts the TPC cluster finder.

# Data compression



- Fast cluster finder in the C-RORC FPGAs:
  - 216 instances in 36 FPGAs (6 per C-RORC, one for every TPC DDL).
  - VHDL (hard to maintain at this level, possible move to a higher level description...)

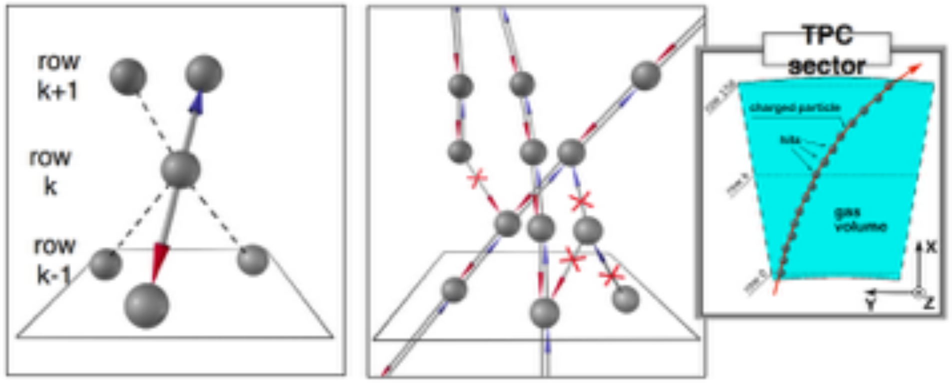


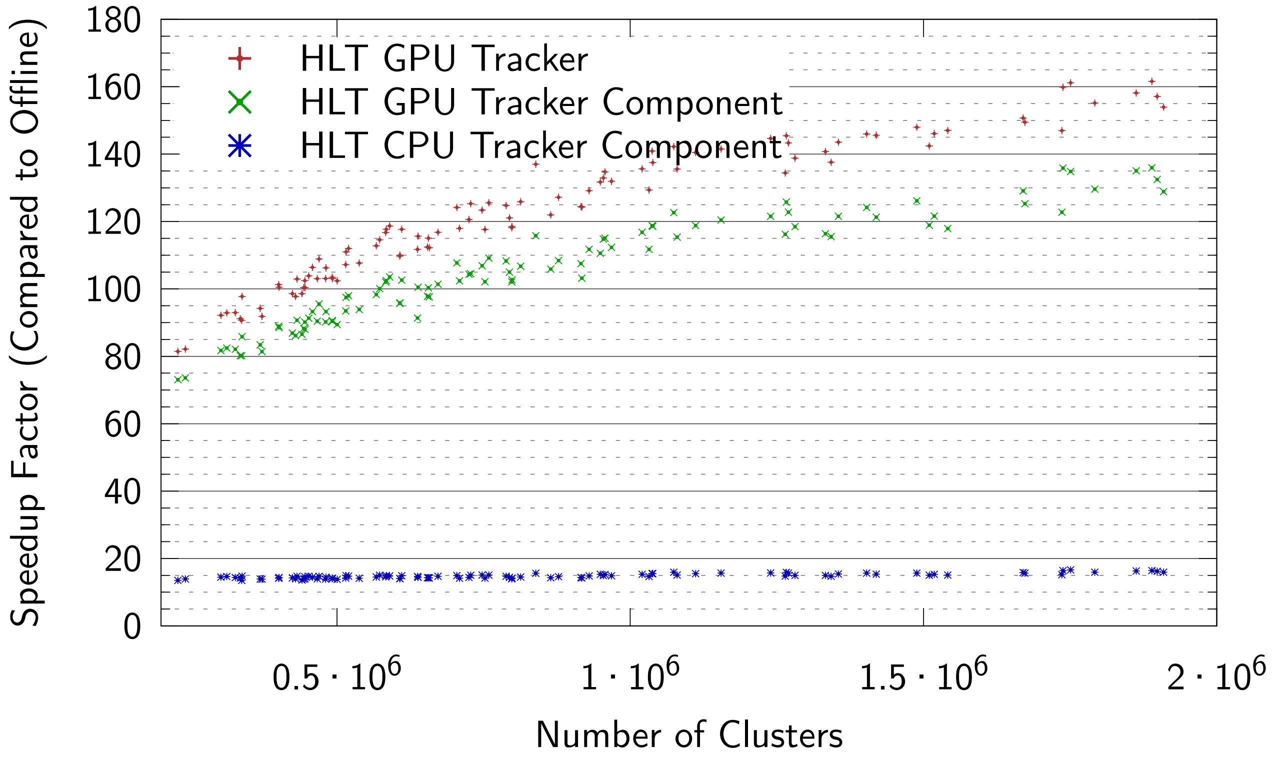
average compression: ~4.3

- Cluster finder output is ordered -> compressible.
  - Data format optimisation.
  - Huffman encoding.

# Online reconstruction

- GPU based cellular automaton track finder.
    - on new farm: OpenCL (AMD GPUs).
    - also a CUDA version (used in Run 1).
    - CPU version (x86 + OpenMP option).
- Same source code for all versions - see talk by D.Rohr.





- factor 10 speed-up wrt. the pure CPU version (or: 1GPU+3CPUs ~ 27CPUs).

# Calibration

- Run 1 calibration: 2 pass scheme:
  1. Calibrate the TPC tracking.
  2. Calibrate detectors that rely on (calibrated) TPC tracking.

- Run 2: Move (at least) the TPC calibration to online.
  - We already have online tracking…

- No changes to the proven HLT data transport framework.
- Has to run at the component level in a wrapper (see talk by D.Rohr).
- Use the EXISTING offline software (with minimal changes).
  - Common interface for online and offline data structures:
  - Same code runs online and offline.
  - Performance optimisations where necessary.

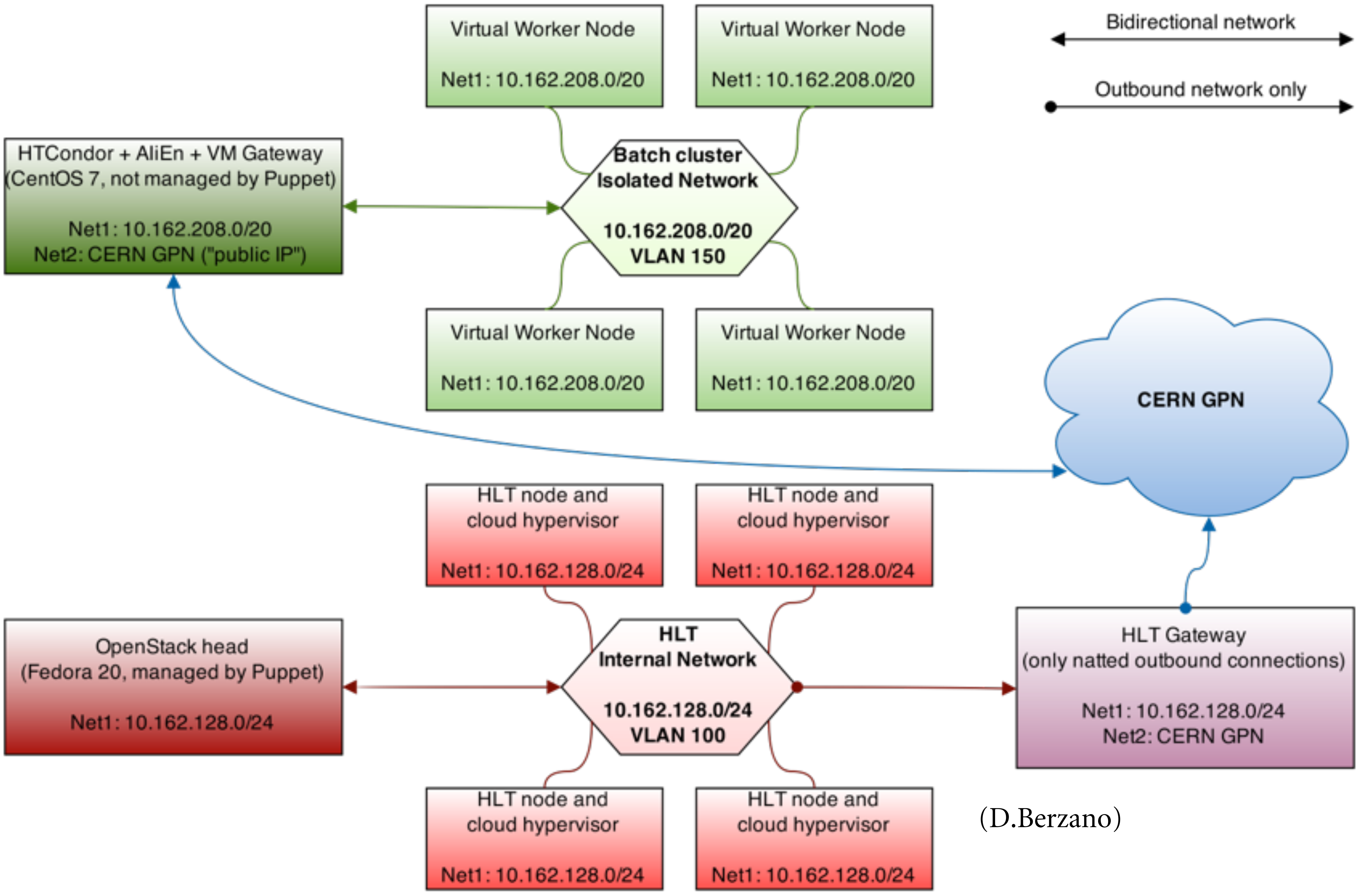
# Opportunistic use for "offline"

- When resources available: technical stops, longer shutdowns
  - ➡ operate as a GRID site.
- OpenStack cloud.
  - Batch system(s): AliEn + HTCondor + elastiq.

- Separation from the online system:
  - VLANs.
  - VMs have dedicated disks (1 SATA SSD per node).

- Under full control of HLT.
- Only centrally managed jobs - no random user jobs.
  - Security.
  - Networking constraints (shared network@Point 2).
- Fully automated config/deployment (Puppet).

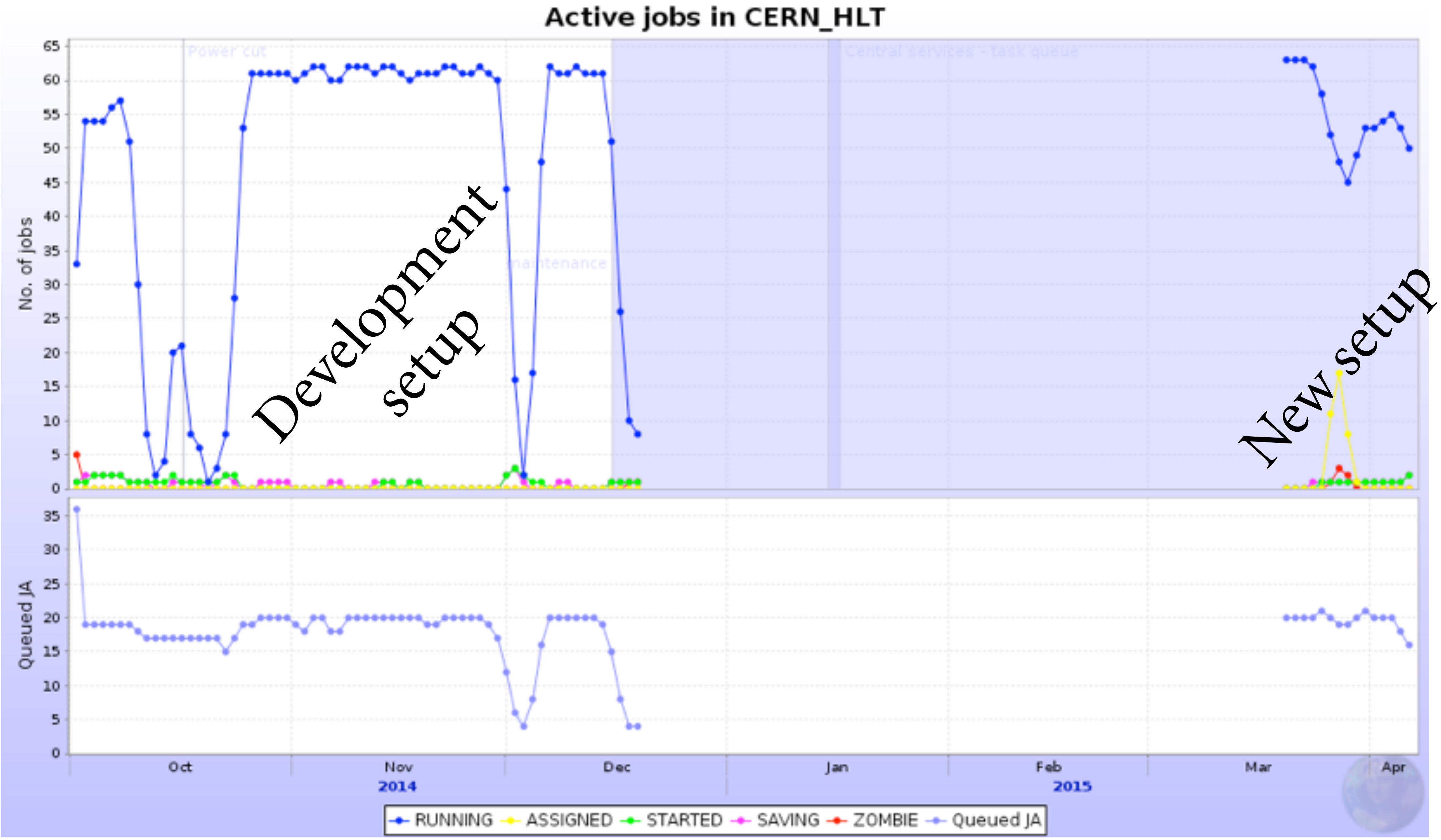
# Openstack setup



(D.Berzano)

- Network separation between the HLT and Openstack.

# CERN_HLT Grid site



- First setup tested last year (development cluster).
- Production grade setup running now (development cluster).
- Use on the production system pending.

# Outlook

- New HLT farm installed @Point 2 and being deployed.
- New use cases:
  - Online calibration.
  - Opportunistic use as a Grid site.

- Future:
  - Continuous development with an eye on the O2 system.
    - GPU tracking.
    - Hardware accelerated cluster finding/compression.