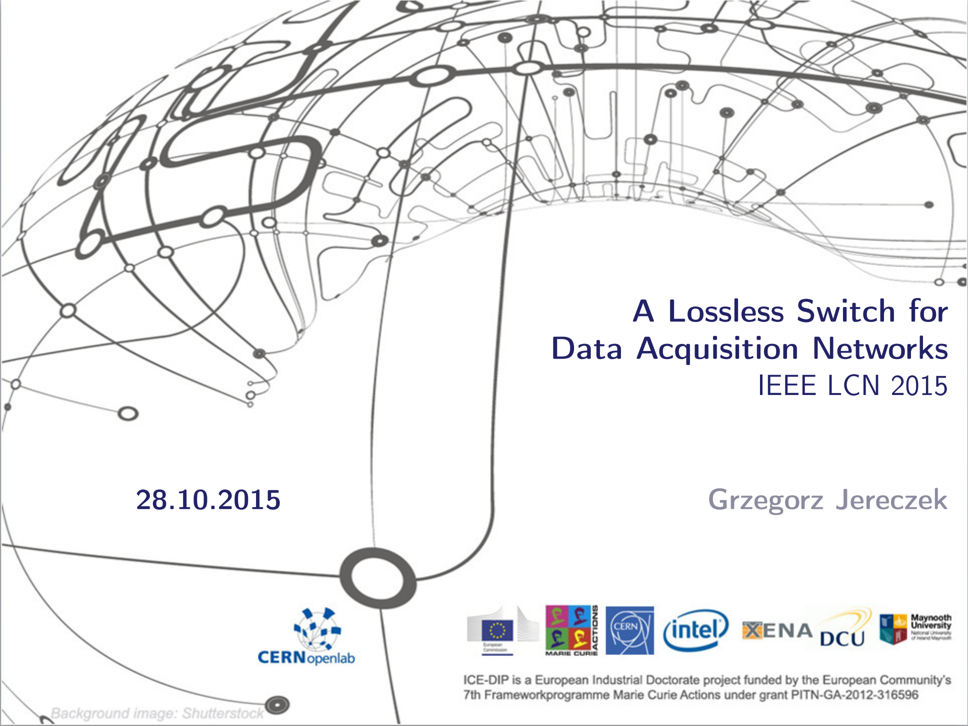


A Lossless Switch for Data Acquisition Networks

IEEE LCN 2015

28.10.2015

Grzegorz Jereczek

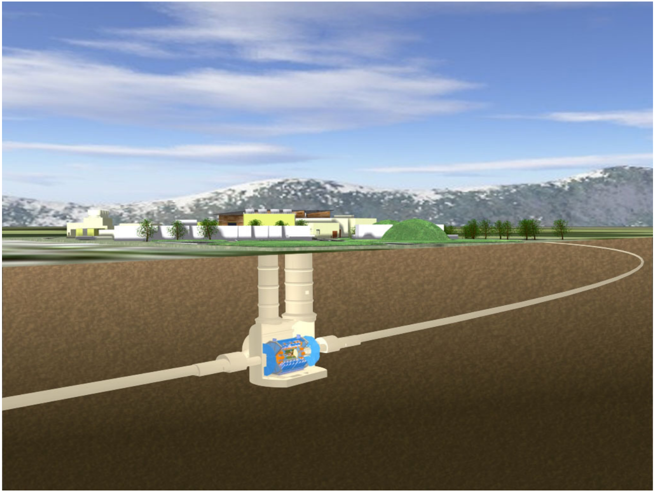
ICE-DIP is a European Industrial Doctorate project funded by the European Community's 7th Frameworkprogramme Marie Curie Actions under grant PITN-GA-2012-316596

Background image: Shutterstock

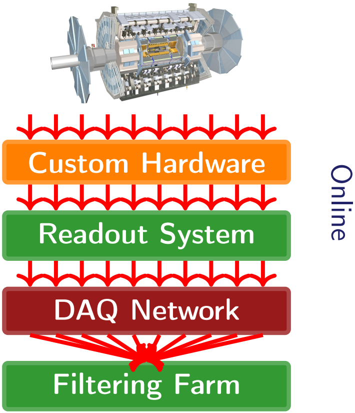
Outline

1. Introduction to data acquisition networks
2. The TCP incast pathology
3. A lossless switch for data acquisition networks
4. Evaluation
5. Conclusions and outlook

Introduction to data acquisition networks



Delivering data to the online filtering farm
Data acquisition (DAQ) networks



Each filtering node requires data fragments from many readout nodes (ROS).

Bursty nature of the traffic and many-to-one communication pattern are a challenge for the network.

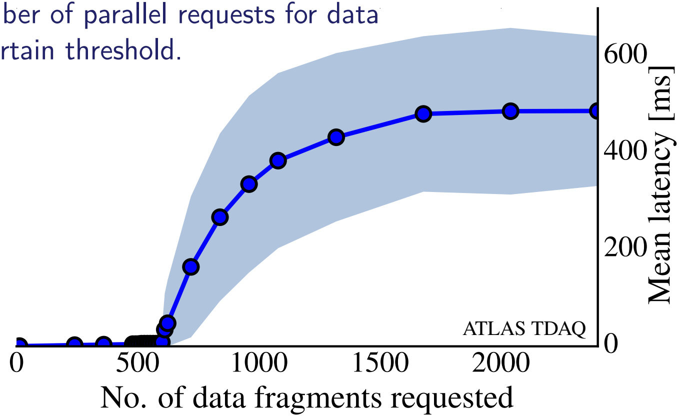
What can be done with commodity TCP/IP?

Data collection time

Critical to the performance of the entire system.

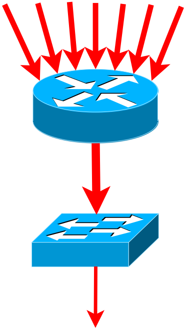
Must be kept under control with low jitter in order to sustain the required long-term throughput.

Increasing the number of parallel requests for data scales only till a certain threshold.



The TCP incast pathology

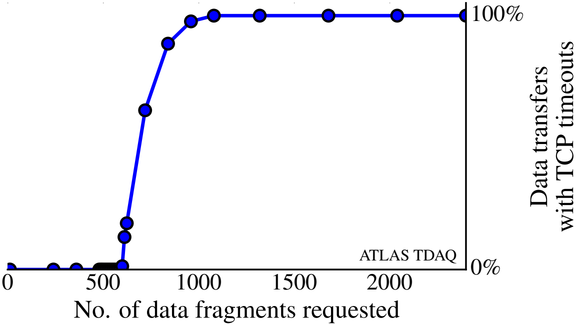
TCP timeouts result in throughput collapse



Switches with small packet buffers drop packets.

TCP waits 200ms for a timeout, flows too small to trigger fast retransmissions.

Analogous to **TCP incast** in datacenter.



Most of the proposals focus on controlling the traffic injected into the network

State of the art: **Data Center TCP (DCTCP)**[1]

- Leverages ECN to keep the switch queues small while maintaining high throughput.
- Fails, if there are so many senders that the packets sent in the first RTT overflow the buffers.

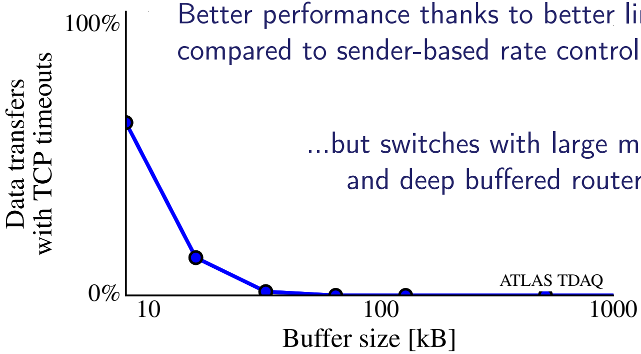
[1] Alizadeh et al., "Data center TCP (DCTCP)". Internet Draft: `https://datatracker.ietf.org/doc/draft-ietf-tcpm-dctcp/`

Prodiving lossless connectivity with large packet buffers

No incast, if bursts size fits within the memory.

Better performance thanks to better link utilization compared to sender-based rate control.

...but switches with large memories are rare and deep buffered routers are expensive. **Alternatives?**



A lossless switch for data acquisition networks

Software switch with packet buffers in DRAM

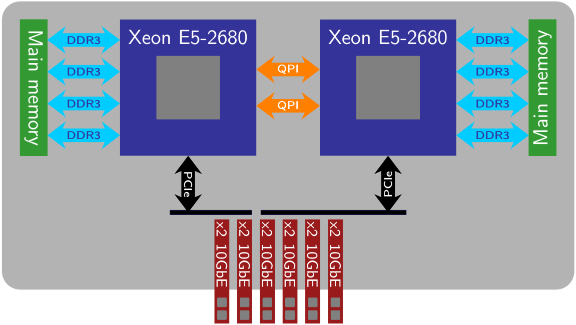
Nearly limitless memory.

Dedicated queuing
to avoid bufferbloat.

Data acquisition networks:

- Throughput-oriented,
- Often based on large packets,
- Relatively small.

Potential limitations do not hold.



DAQ network based on software switches?

Prototype: **12 x 10GbE ports**

DPDK framework for building fast packet processing applications (`http://dpdk.org/`)

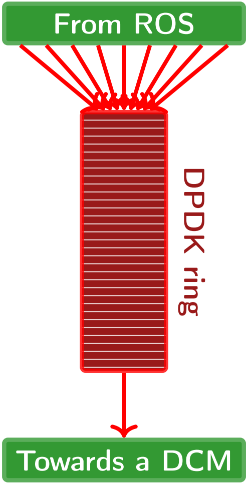
The x86 DPDK-based switching application

Dedicated queue for each incast-sensitive destination data collector (DCM).

Packets queued in the DPDK's rings.

Single ring dedicated to single data collector.

Rate limitation can be applied
to prevent incast in subsequent hops.

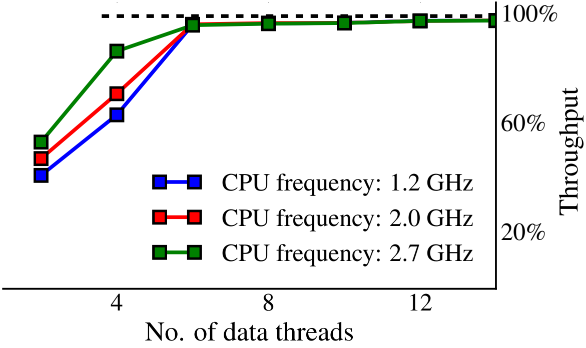


Evaluation

Evaluating the offered bandwidth
All-to-all traffic: 12 ROSes and 144 DCMs

97% of theoretical throughput with 6 CPU cores @1.2GHz.

Utilizing full bidirectional bandwidth of 120Gbps.

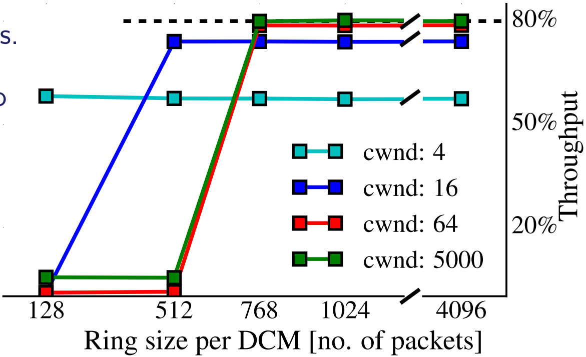


Applying rate limitation
All-to-all traffic: 12 ROSes and 144 DCMs

Rate limit of 0.78Gbps for each destination DCM (990 pkts / 11 flows).
Packet buffer: 1.12GiB (144 rings x 4096 pkts).

Emulating incast avoidance in subsequent network nodes.

Ring size can be adjusted to match the total burst size towards a DCM.



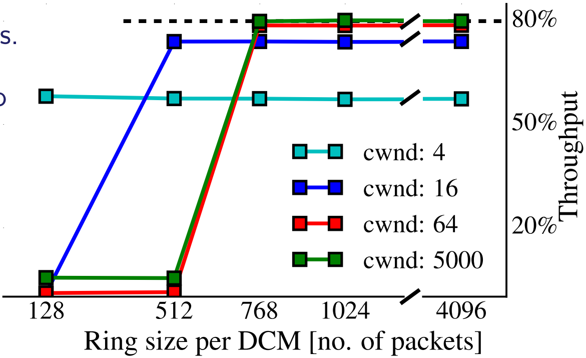
Applying rate limitation
All-to-all traffic: 12 ROSes and 144 DCMs

Rate limit of 0.78Gbps for each destination DCM (990 pkts / 11 flows).
Packet buffer: 1.12GiB (144 rings x 4096 pkts).

Emulating incast avoidance in subsequent network nodes.

Ring size can be adjusted to match the total burst size towards a DCM.

Better performance without packet injection control (static congestion window).



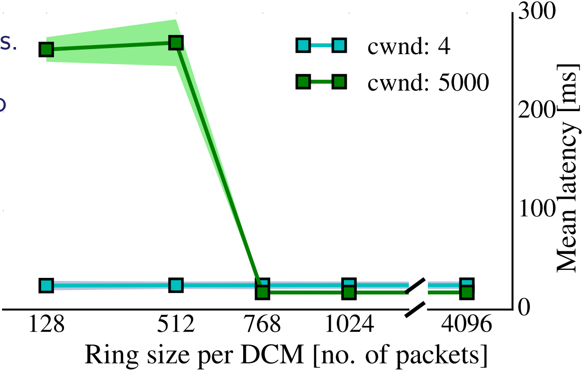
Applying rate limitation
All-to-all traffic: 12 ROSes and 144 DCMs

Rate limit of 0.78Gbps for each destination DCM (990 pkts / 11 flows).
Packet buffer: 1.12GiB (144 rings x 4096 pkts).

Emulating incast avoidance in subsequent network nodes.

Ring size can be adjusted to match the total burst size towards a DCM.

Better performance without packet injection control (static congestion window).



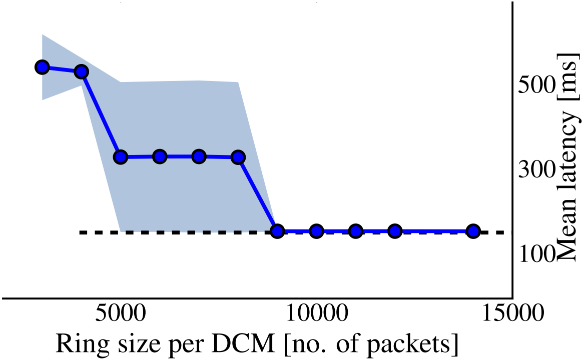
Evaluating buffering capabilities
All-to-one traffic: 110 ROSes and 1 DCM

Rate limit of 0.78Gbps (9790 pkts / 110 flows).
Packet buffer: 27.3MiB (1 ring x 14000 pkts).

Increased burstiness with a single DCM.

Incast for ring sizes below 9000 packets.

No incast with expected mean latency, and no jitter otherwise.



Conclusions and outlook

Trying to prevent incast congestion in DAQ

DRAM memory provides large enough
and cheap packet buffers.

Dedicated queueing to optimize the entire network.

Prototype offers **lossless operation** and **120Gbps bandwidth**
for DAQ-specific network traffic.

Could we build the entire DAQ network with software switches?

Bandwidth-wise, the prototype provides figures comparable to the requirements of the existing system.

The architecture needs to scale
for the future LHC upgrades...

...and provide the required port density.

Configuration and management aspects
are not less important.

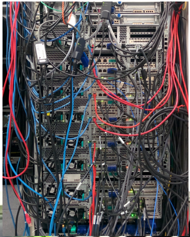
Potential topology with a mixture of ToR and dedicated software switches

ToR switches to provide port density.

Software switches to provide packet buffers and configured to mitigate incast: **lossless operation**.

Configuration and management with SDN: **Open vSwitch**, **OpenFlow**, **OVSDB**.

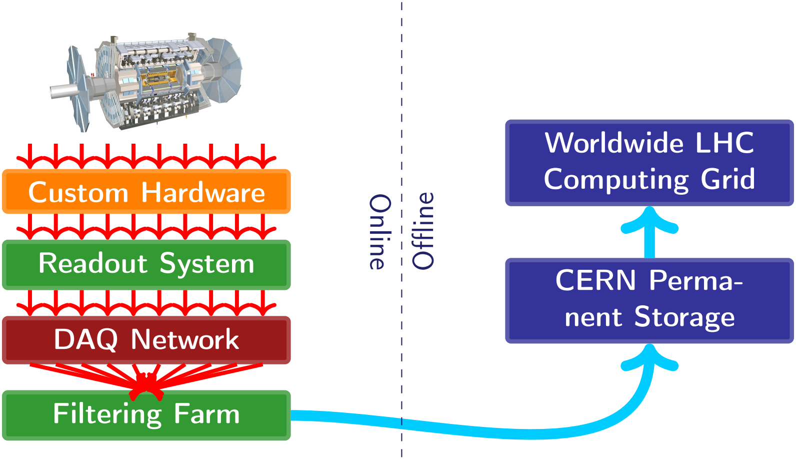
Questions?



B4
B1
21
2
PC H
PC G
PC E
PC D
AIPC P/N 571540002
1127
Intel Technology

Backup

Data flow of the ATLAS experiment at CERN



Custom Hardware

Readout System

DAQ Network

Filtering Farm

Online

Offline

Worldwide LHC Computing Grid

CERN Permanent Storage

Reconstruct, analyse and select complex events in real time.

Ways to approach TCP incast

$$\mathbf{BDP} + \mathbf{BufferSize} < \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{wnd_i}$$

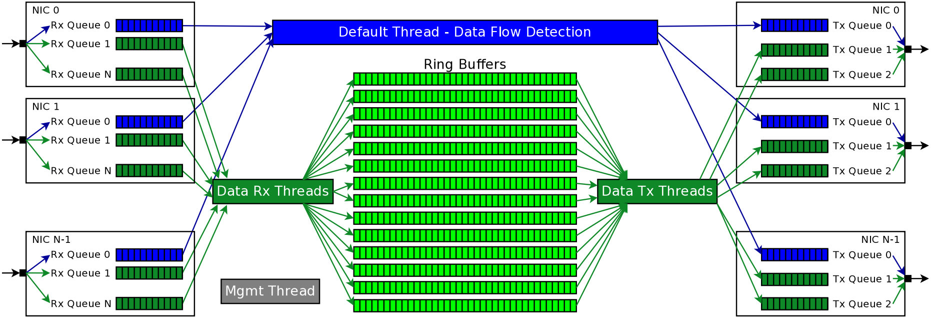
Increase the link speeds.

Extend the buffers.

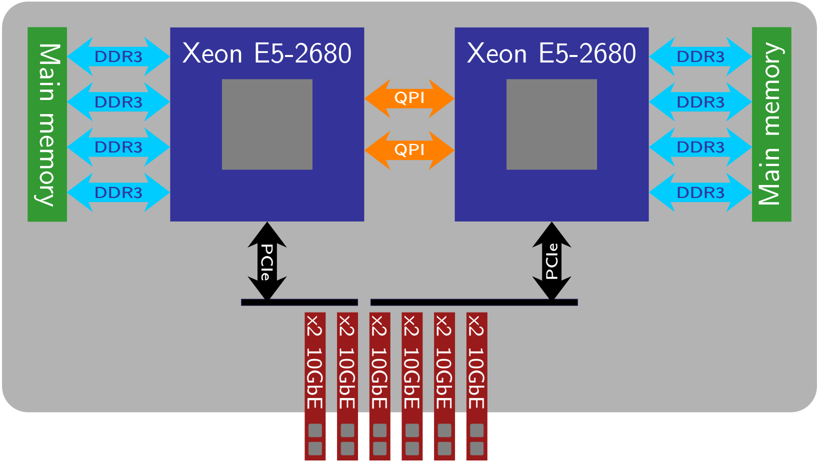
Keep the global window under control at the:

- link layer,
- transport layer,
- application layer.

The x86 DPDK-based switching application



A prototype of a software switch with packet buffers in DRAM memory



Performance was always the challenge What has changed?

Recent developments in commodity servers:

- Integrated memory controller,
- Direct PCIe lanes to CPU,
- Memory: 340 Gbps with DDR-10600 and 4 channels per CPU,
- PCIe: 32 Gbps Gen2 and 63 Gbps Gen3 (x8),
- Direct Memory Access (DMA) and Direct Cache Access (DCA),
- Modern NICs features, e.g. RSS, offloads.

Near real-time kernel configuration:

- CPU cores isolation,
- Tickless kernel.

The raise of fast packet processing frameworks, e.g.:

- **DPDK**, PF_RING, netmap, Snabb Switch.

Overcoming the limits of ECN- and QCN-based solutions in hardware

Limited number of traffic classes in traditional switches and routers.

Software switching offers scalability in the number of queues.

Dedicated design with separate queue for each data collector:

- Better fairness,
- No bufferbloat,
- Preventing incast with appropriate queue size,
- Preventing incast in subsequent network stages with rate limitation.

Evaluation setup

Device under test

Single instance of the lossless software switch.

Xeon-based commodity server with 12x10GbE ports.

In theory, 120Gbps of offered bandwidth.

Traffic generation

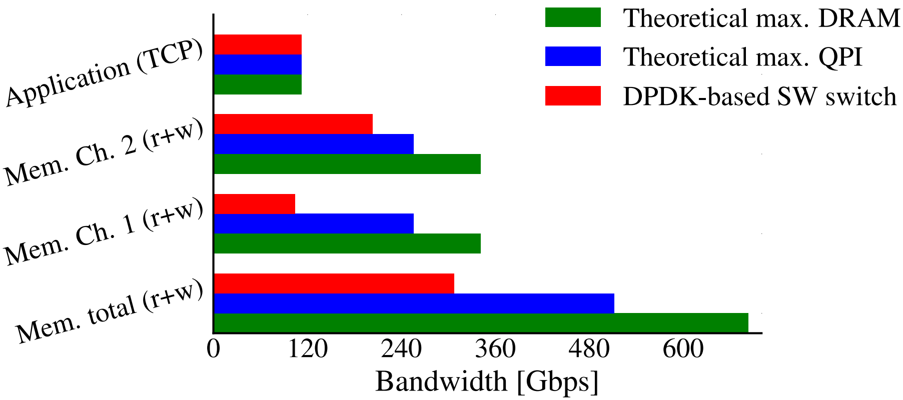
ATLAS DAQ/HLT software in emulation mode.

Data providers (ROS) and collectors (DCM) running on all 12 hosts connected to the switch.

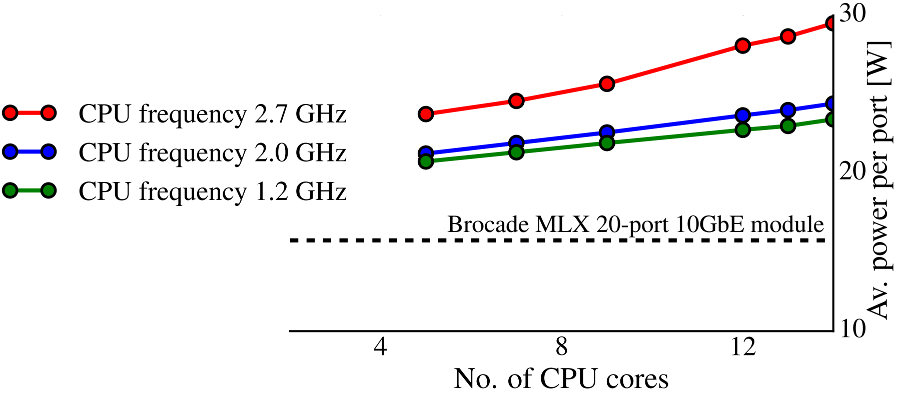
1500B MTU.

TCP congestion control disabled.

Memory bandwidth usage



Power consumption


CPU frequency 2.7 GHz
CPU frequency 2.0 GHz
CPU frequency 1.2 GHz
Brocade MLX 20-port 10GbE module
Av. power per port [W]
No. of CPU cores