White Rabbit

Ethernet-based solution for sub-ns synchronization and deterministic, reliable data delivery

Maciej Lipiński

on behalf of White Rabbit Team

Hardware and Timing Section @ CERN
Institute of Electronic Systems @ Warsaw University of Technology

15 July 2013
IEEE Plenary Meeting Genève

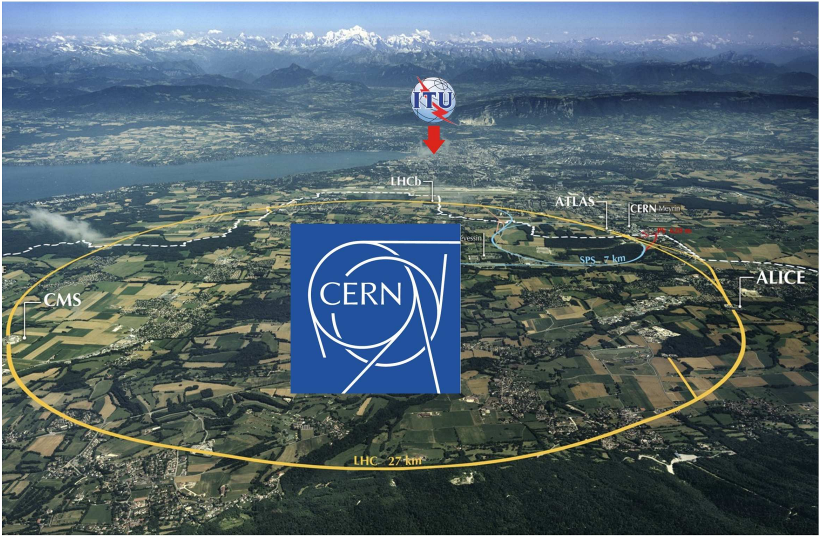
Outline

1. Introduction
2. CERN Control & Timing
3. WR Network
4. Time Distribution
   - Timing demo
5. Data Distribution
   - Redundancy demo
6. WR @ CERN
7. Summary

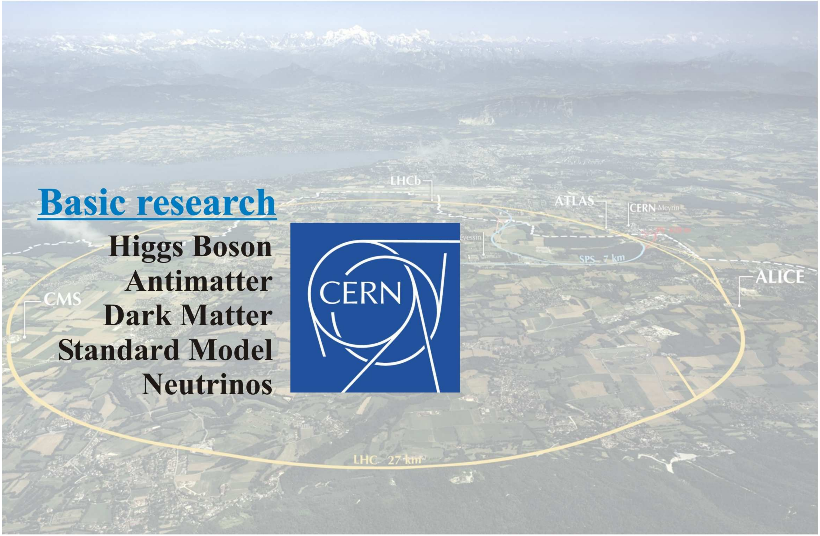
Outline

1. Introduction
2. CERN Control & Timing
3. WR Network
4. Time Distribution
   - Timing demo
5. Data Distribution
   - Redundancy demo
6. WR @ CERN
7. Summary

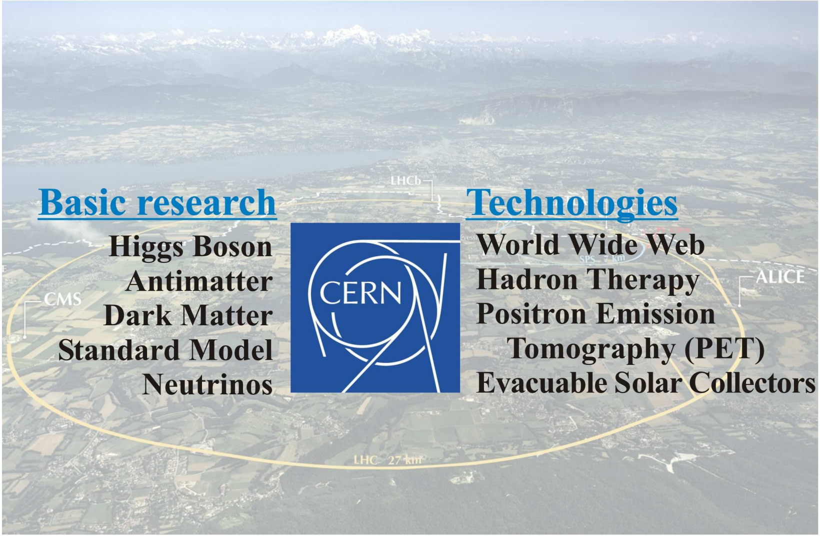
CERN



CERN


Basic research
Higgs Boson
Antimatter
Dark Matter
Standard Model
Neutrinos
CERN
LHCb
ATLAS
CERN Meyrin
CMS
ALICE
SPS 7 km
LHC 27 km

CERN



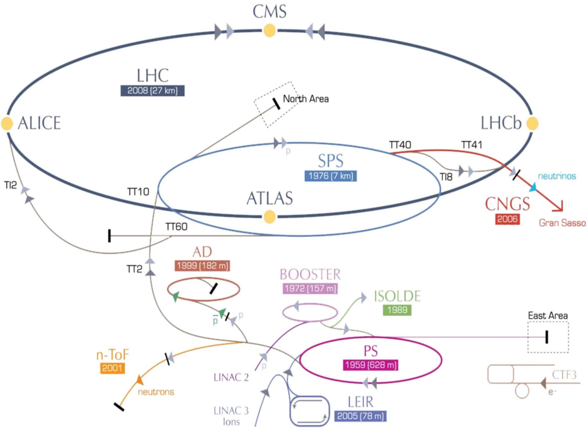
Basic research

Higgs Boson
Antimatter
Dark Matter
Standard Model
Neutrinos

Technologies

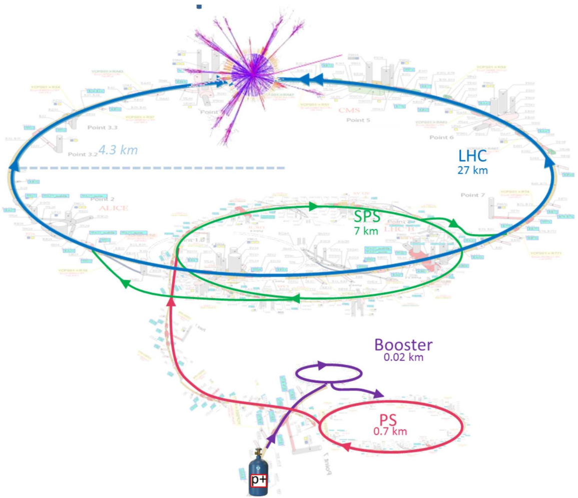
World Wide Web
Hadron Therapy
Positron Emission Tomography (PET)
Evacuable Solar Collectors

CERN Accelerator Complex

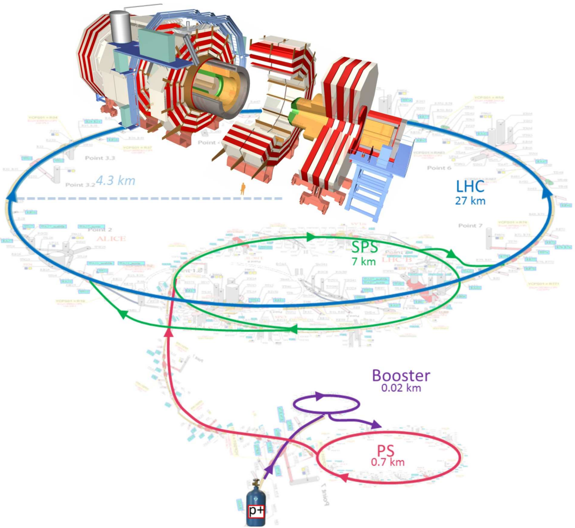


CERN Accelerator Complex

CERN Accelerator Complex



CERN Accelerator Complex



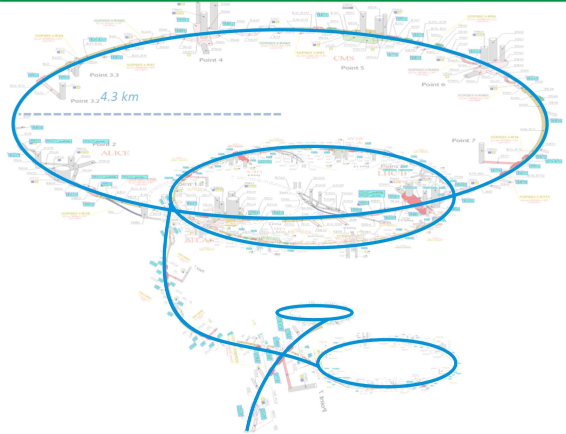
Beams – Controls – Hardware & Timing

Beams – Controls – Hardware & Timing

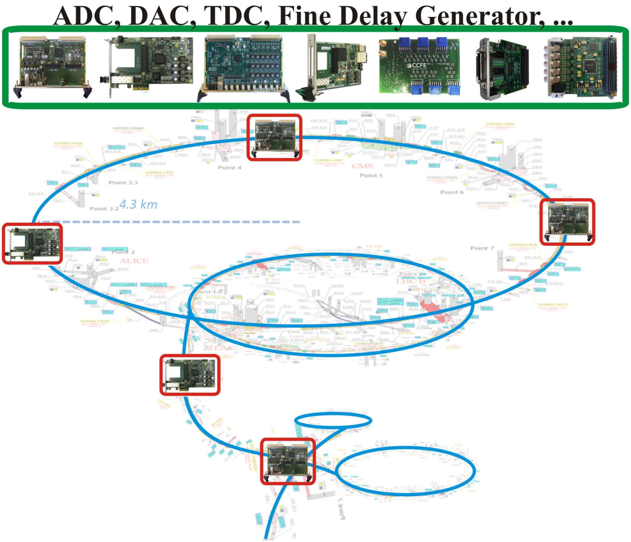
ADC, DAC, TDC, Fine Delay Generator, ...

Beams – Controls – Hardware & Timing

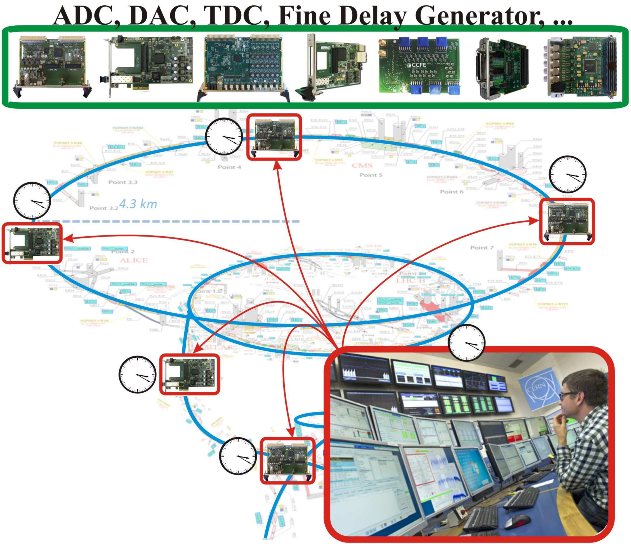
ADC, DAC, TDC, Fine Delay Generator, ...



Beams – Controls – Hardware & Timing



Beams – Controls – Hardware & Timing

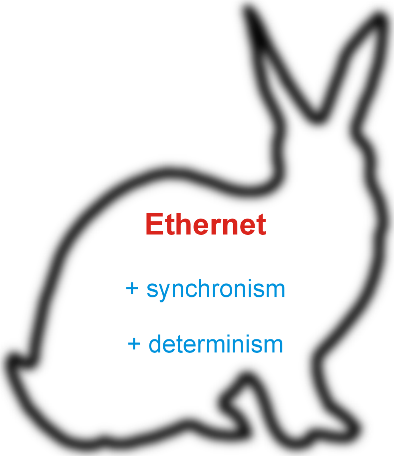


What is White Rabbit?

- Renovation of accelerator's control and timing
- Based on well-known technologies
- Open Hardware and Open Software
- International collaboration

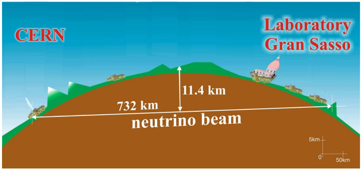
White Rabbit features

- standard-compatible
- sub-ns accuracy
- tens-ps precision
- upper-bound low-latency
- white-box simulation & analysis
- high reliability
- tens-km span
- thousands-nodes systems



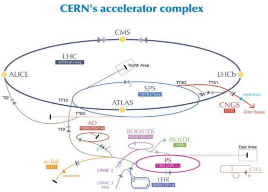
White Rabbit applications

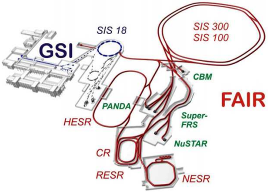
- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso



White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI





White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI
  - HiSCORE: Gamma&Cosmic-Ray experiment (Tunka, Siberia)

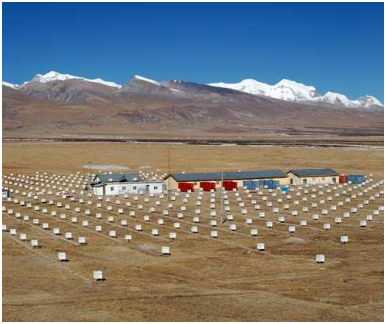
DESY

Karlsruhe Institute of Technology

> Institute for Nuclear Research of the Russian Academy of Sciences
> Moscow State University
> Irkutsk State University

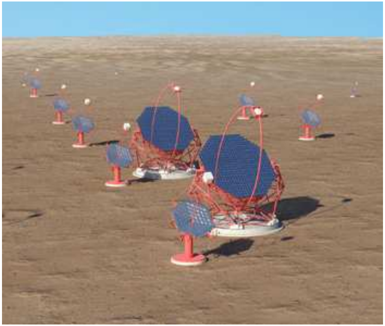
White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI
  - HiSCORE: Gamma&Cosmic-Ray experiment (Tunka, Siberia)
  - The Large High Altitude Air Shower Observatory (China)



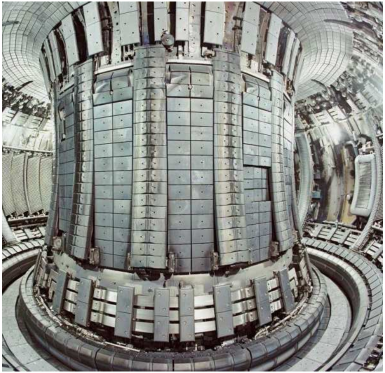
White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI
  - HiSCORE: Gamma&Cosmic-Ray experiment (Tunka, Siberia)
  - The Large High Altitude Air Shower Observatory (China)
- Potential applications:
  - Cherenkov Telescope Array



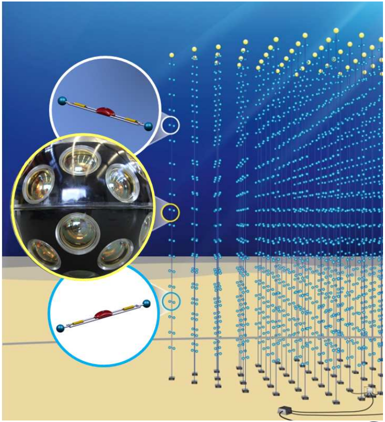
White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI
  - HiSCORE: Gamma&Cosmic-Ray experiment (Tunka, Siberia)
  - The Large High Altitude Air Shower Observatory (China)
- Potential applications:
  - Cherenkov Telescope Array
  - International Thermonuclear Experimental Reactor (ITER)



White Rabbit applications

- Deployed for time distribution:
  - CERN Neutrinos to Gran Sasso
- Future applications:
  - CERN and GSI
  - HiSCORE: Gamma&Cosmic-Ray experiment (Tunka, Siberia)
  - The Large High Altitude Air Shower Observatory (China)
- Potential applications:
  - Cherenkov Telescope Array
  - International Thermonuclear Experimental Reactor (ITER)
  - European deep-sea research infrastructure (KM3NET)



Outline

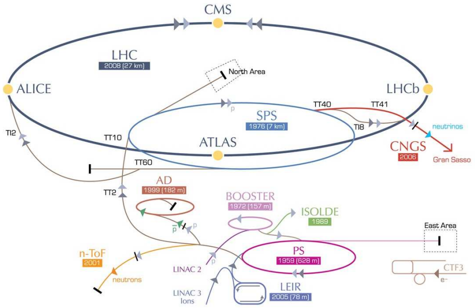
1 Introduction
2 CERN Control & Timing
3 WR Network
4 Time Distribution
- Timing demo

5 Data Distribution
- Redundancy demo

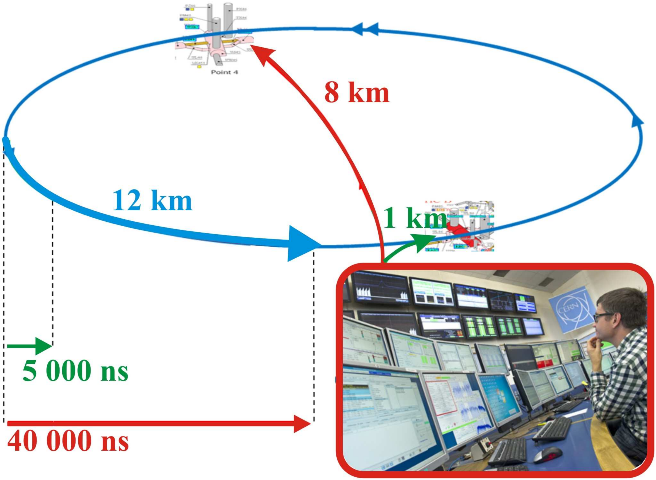
6 WR @ CERN
7 Summary

CERN Control and Timing System

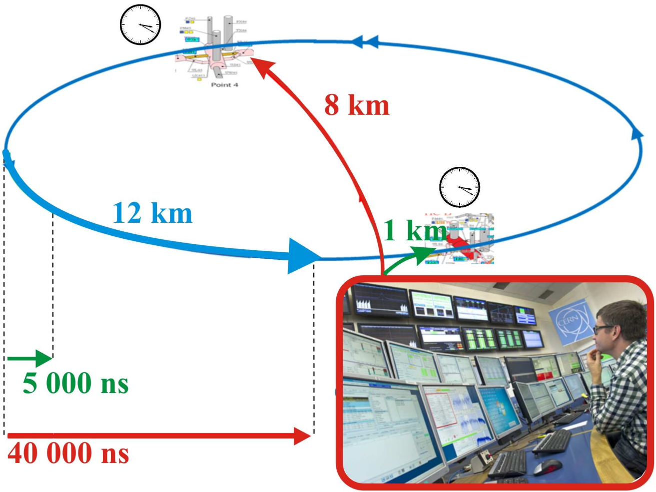
- 6 accelerators including LHC: 27km
- A huge real-time distributed system
- Thousands of devices



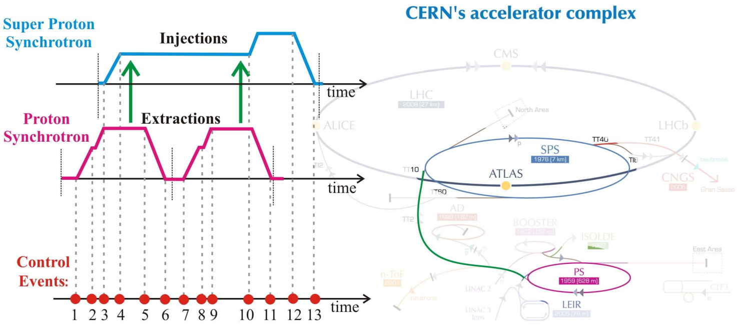
Controlling accelerators



Controlling accelerators

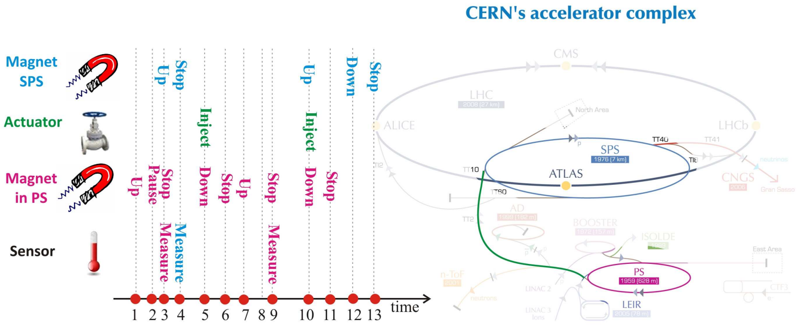


CERN Control System – event distribution (1)



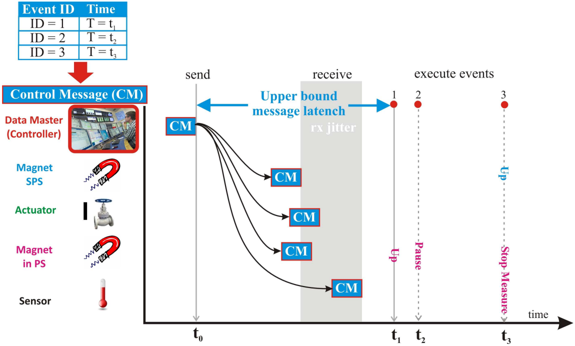
- **Events** – messages which trigger actions
- Each event is identified by an **ID**

CERN Control System – event distribution (2)



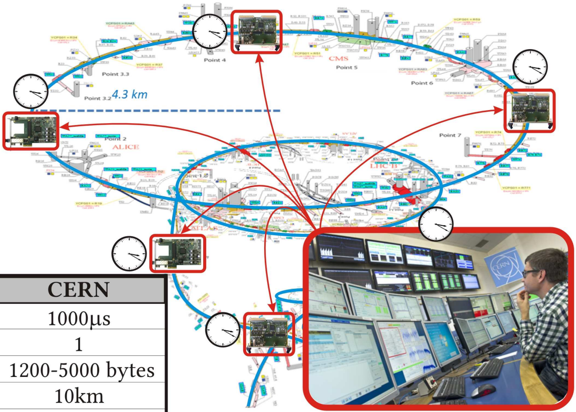
- Devices are subscribed to events
- Each device "knows" what to do on a particular event

CERN Control System – event distribution (3)



- Each event (ID) has a trigger time associated
- A set of events is sent as a single **Control Message (CM)**
- CM is broadcast to all the end devices (nodes)
- CM is sent in advance (**upper-bound message latency**)

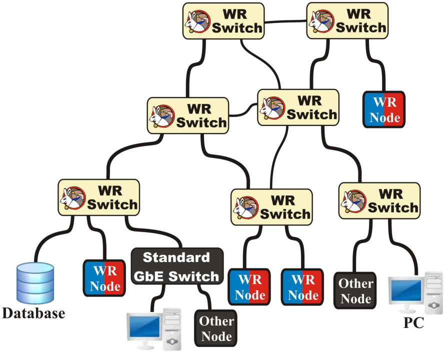
CERN Control & Timing Network – requirements



Requirement	CERN
Max latency	1000µs
CMs lost per year	1
CM size	1200-5000 bytes
Network span	10km
Accuracy	up to 1ns

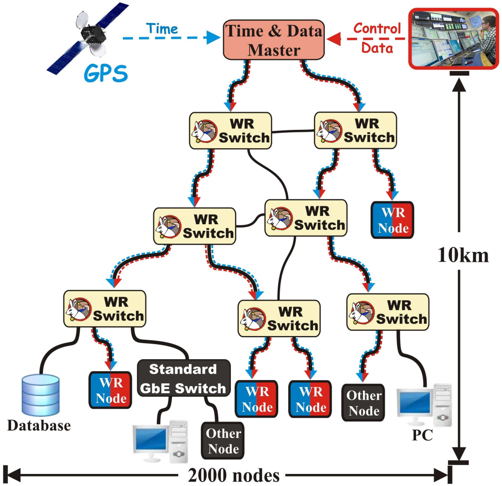
White Rabbit Network – Ethernet-based

- Standard Ethernet network
- Few thousands nodes
- Bandwidth: 1 Gbps
- WR Switch: 18 ports
- Non-WR Devices
- Ethernet features (VLAN) & protocols (SNMP)



White Rabbit Network – Ethernet-based

- High accuracy/precision synchronization
- Deterministic, reliable and low-latency Control Data delivery

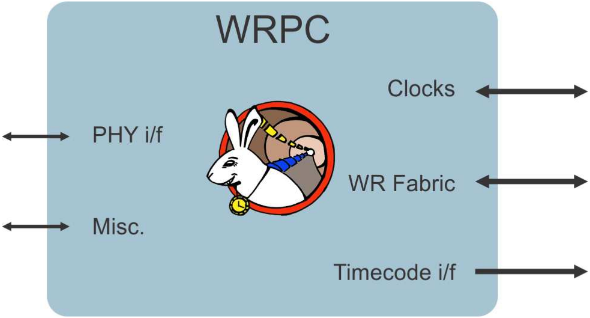


White Rabbit Switch

- Central element of WR network
- Original design optimized for timing, designed from scratch
- 18 ports
- 1000BASE-BX10 SFPs: up to 10 km, single-mode fiber
- Open design (H/W and S/W)

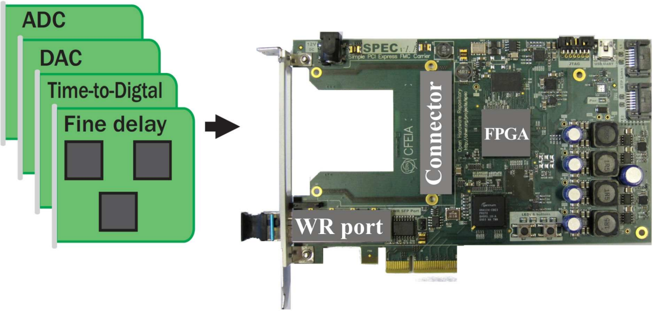
White Rabbit Node

- Ethernet MAC with White Rabbit
  - Open IP Core
  - Easily integrated into custom FPGA-based designs



White Rabbit Node

- Ethernet MAC with White Rabbit
  - Open IP Core
  - Easily integrated into custom FPGA-based designs
- WR Node: universal carrier board



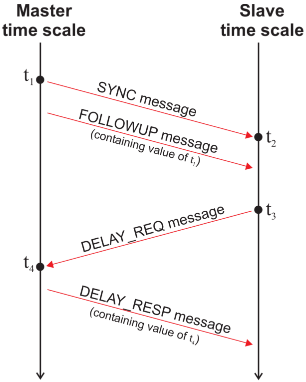
Outline

1. Introduction
2. CERN Control & Timing
3. WR Network
4. Time Distribution
   - Timing demo
5. Data Distribution
   - Redundancy demo
6. WR @ CERN
7. Summary

Time Distribution in White Rabbit Network

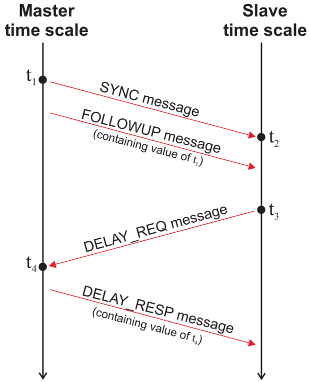
- Synchronization with **sub-ns** accuracy **tens-ps** precision
- Combination of
  - Precision Time Protocol (**IEEE1588**) synchronization
  - Layer 1 syntonization
  - Digital Dual-Mixer Time Difference (**DDMTD**) phase detection

Precision Time Protocol (IEEE1588)



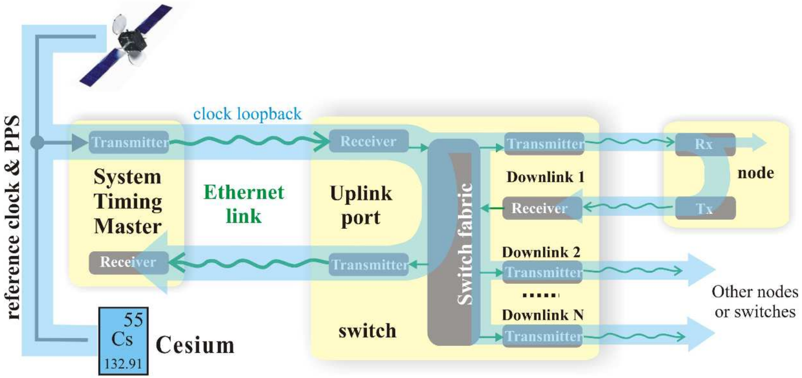
- Simple calculations:
  - link *delay*$_{ms}$: $\delta_{ms} = \frac{(t_4 - t_1) - (t_3 - t_2)}{2}$
  - clock *offset*$_{ms} = t_2 - t_1 + \delta_{ms}$
- Assumes medium symmetry

Precision Time Protocol (IEEE1588)



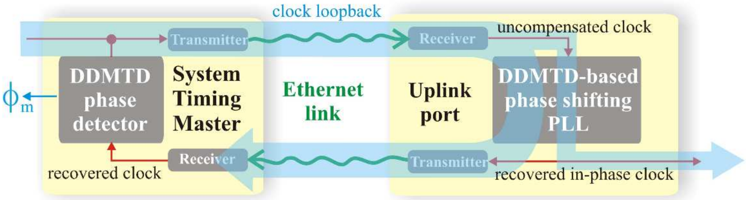
- Simple calculations:
  - link *delay*$_{ms}$: $\delta_{ms} = \frac{(t_4 - t_1) - (t_3 - t_2)}{2}$
  - clock *offset*$_{ms} = t_2 - t_1 + \delta_{ms}$
- Assumes medium symmetry
- Disadvantages
  - all nodes have free-running oscillators
  - frequency drift compensation vs. message exchange traffic

Layer 1 Syntonization

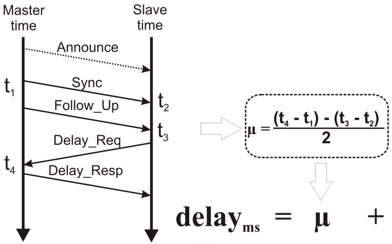


Phase Tracking (DDMTD)

- Monitor phase of bounced-back clock
- Enhance PTP timestamps with phase measurement
- Phase-locked loop in the slave follows the phase changes



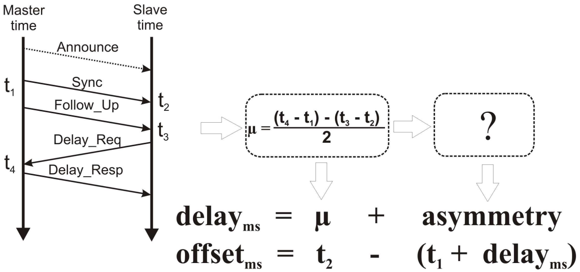
Link Delay Model



$$\text{delay}_{ms} = \mu + \text{asymmetry}$$

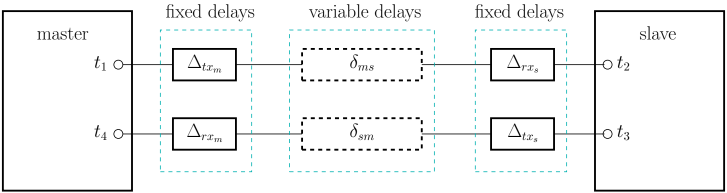
$$\text{offset}_{ms} = t_2 - (t_1 + \text{delay}_{ms})$$

Link Delay Model



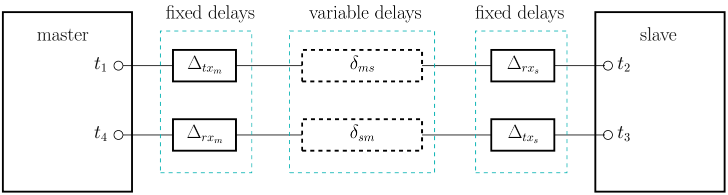
Link Delay Model

$$delay_{ms} = \Delta_{tx_m} + \delta_{ms} + \Delta_{rx_s}$$
$$delay_{sm} = \Delta_{tx_s} + \delta_{sm} + \Delta_{rx_m}$$



Link Delay Model

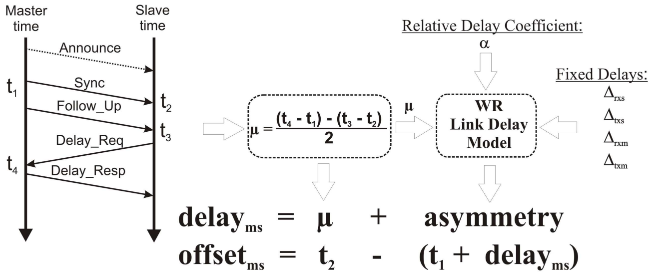
$$delay_{ms} = \Delta_{tx_m} + \delta_{ms} + \Delta_{rx_s}$$
$$delay_{sm} = \Delta_{tx_s} + \delta_{sm} + \Delta_{rx_m}$$



Relative Delay Coefficient (α)
for 1000base-X over a Single-mode Optical Fibre

$$\delta_{ms} = (1 + \alpha)\,\delta_{sm}$$

Link Delay Model: fiber optic solution

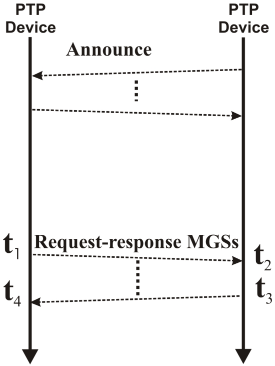


Solution for Ethernet over a Single-mode Optical Fiber

$$asymmetry = \Delta_{tx_m} + \Delta_{rx_s} - \frac{\Delta - \alpha\mu + \alpha\Delta}{2 + \alpha}$$

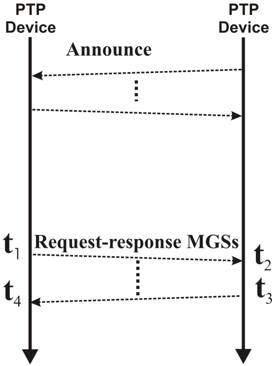
White Rabbit extension to PTP

- White Rabbit requires:
  - WR-specific states
  - Exchange of WR-specific information



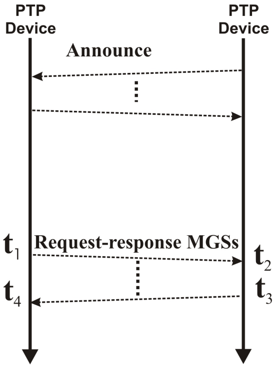
White Rabbit extension to PTP

- White Rabbit requires:
  - WR-specific states
  - Exchange of WR-specific information
- White Rabbit estimates link asymmetry



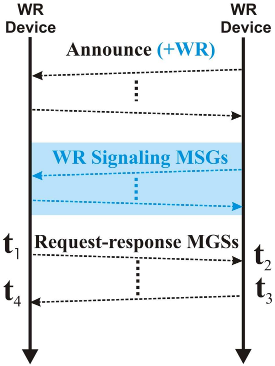
White Rabbit extension to PTP

- White Rabbit requires:
  - WR-specific states
  - Exchange of WR-specific information
- White Rabbit estimates link asymmetry
- WR PTP



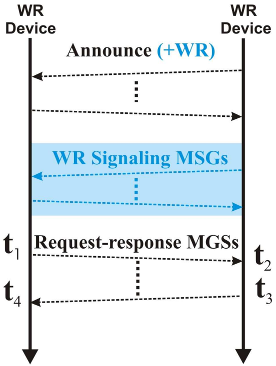
White Rabbit extension to PTP

- White Rabbit requires:
  - WR-specific states
  - Exchange of WR-specific information
- White Rabbit estimates link asymmetry
- WR PTP
  - PTP extensions mechanisms
  - Enhanced precision t_1, t_2, t_3, t_4
  - Correction for asymmetry
  - Interoperability with PTP gear



White Rabbit extension to PTP

- White Rabbit requires:
  - WR-specific states
  - Exchange of WR-specific information
- White Rabbit estimates link asymmetry
- WR PTP
  - PTP extensions mechanisms
  - Enhanced precision t_1, t_2, t_3, t_4
  - Correction for asymmetry
  - Interoperability with PTP gear



ISPCS Plug Fest

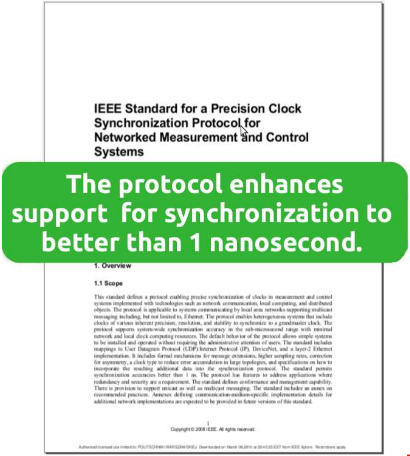
WR: most accurate PTP implementation in the world!

WR Standardization under IEEE1588

- We want to standardize!

WR Standardization under IEEE1588

- We want to standardize!
- Intention by p1588 SG expressed in PAR



IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems

The protocol enhances support for synchronization to better than 1 nanosecond.

1. Overview

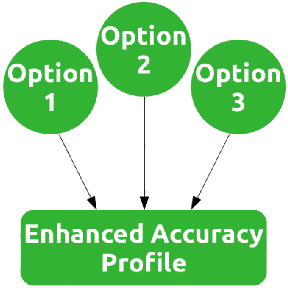
1.1 Scope

This standard defines a protocol enabling precise synchronization of clocks in measurement and control systems implemented with technologies such as network communication, local computing, and distributed objects. The protocol is applicable to systems communicating by local area networks supporting multicast messaging including, but not limited to, Ethernet. The protocol enables heterogeneous systems that include clocks of various inherent precision, resolution, and stability to synchronize to a grandmaster clock. The protocol supports system-wide synchronization accuracy in the sub-microsecond range with minimal network and local clock computing resources. The default behavior of the protocol allows simple systems to be installed and operated without requiring the administrative attention of users. The standard includes mappings to User Datagram Protocol (UDP)/Internet Protocol (IP), DeviceNet, and a layer-2 Ethernet implementation. It includes formal mechanisms for message extensions, higher sampling rates, correction for asymmetry, a clock type to reduce error accumulation in large topologies, and specifications on how to incorporate the resulting additional data into the synchronization protocol. The standard permits synchronization accuracies better than 1 ns. The protocol has features to address applications where redundancy and security are a requirement. The standard defines conformance and management capability. There is provision to support unicast as well as multicast messaging. The standard includes an annex on recommended practices. Annexes defining communication-medium-specific implementation details for additional network implementations are expected to be provided in future versions of this standard.

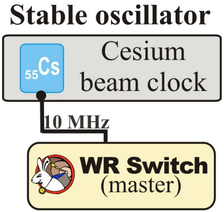
1

WR Standardization under IEEE1588

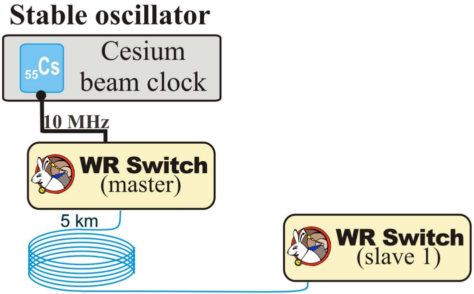
- We want to standardize!
- Intention by p1588 SG expressed in PAR
- Enhanced Accuracy Options / Profile



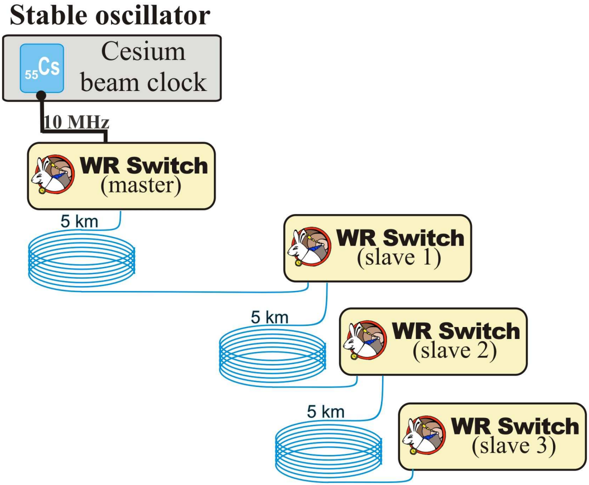
WR synchronization performance



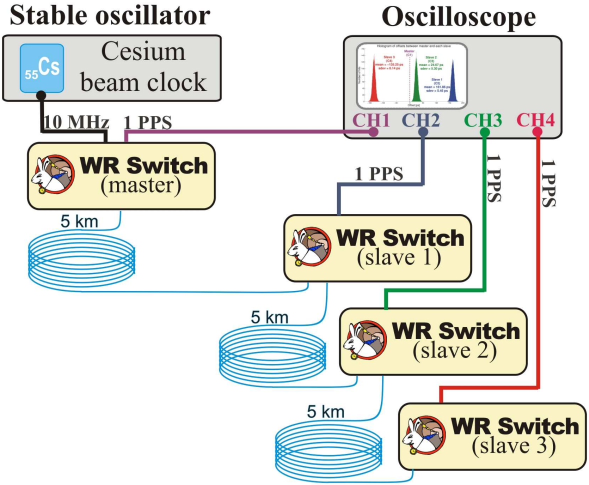
WR synchronization performance



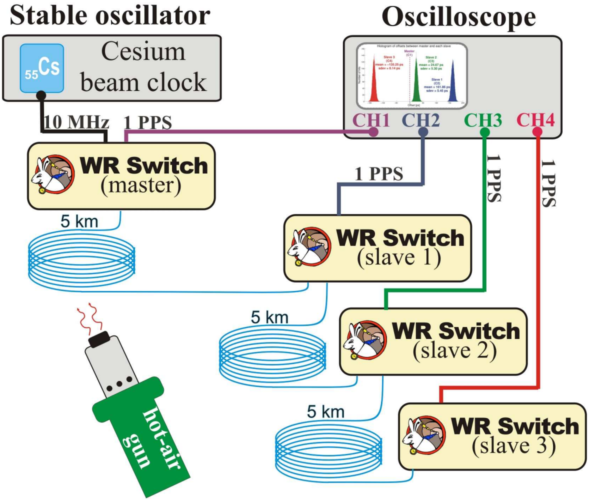
WR synchronization performance



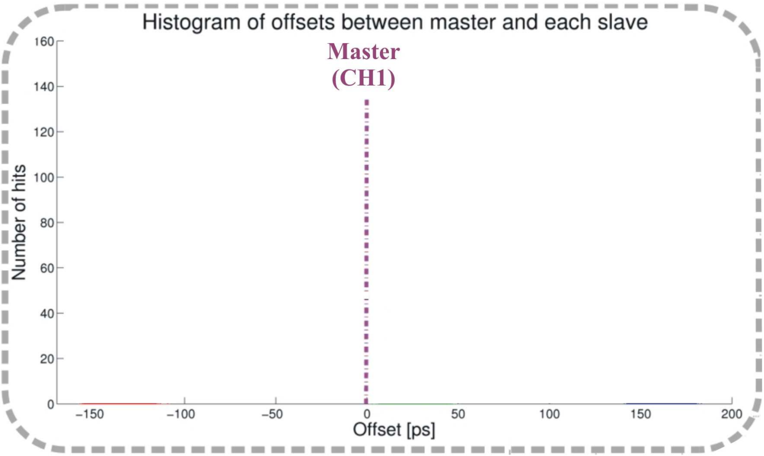
WR synchronization performance



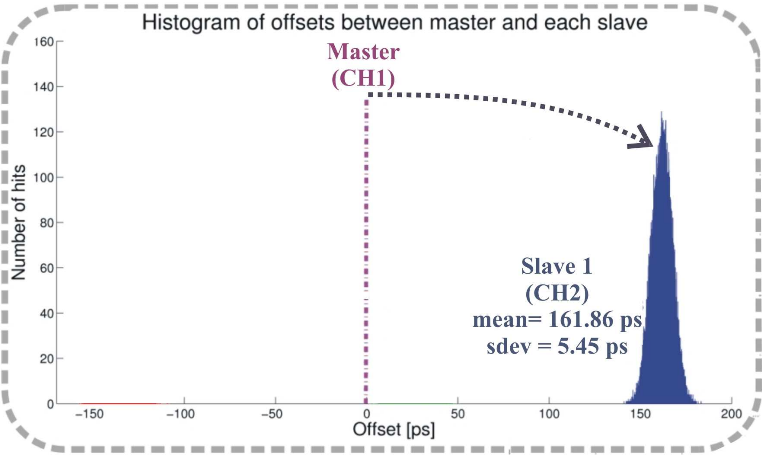
WR synchronization performance



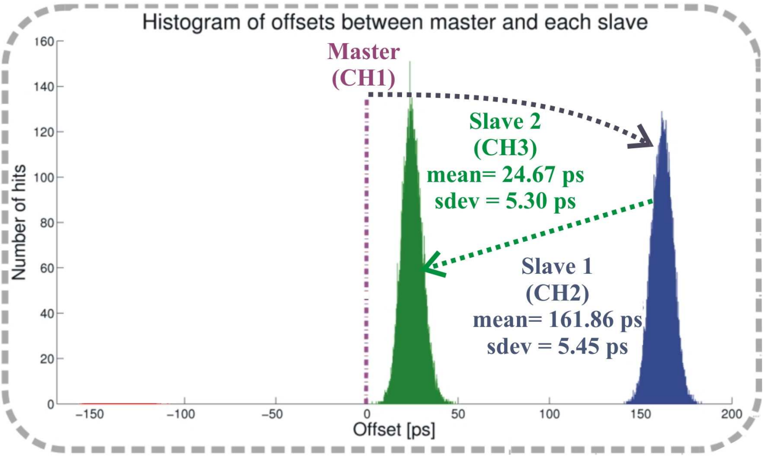
WR synchronization performance



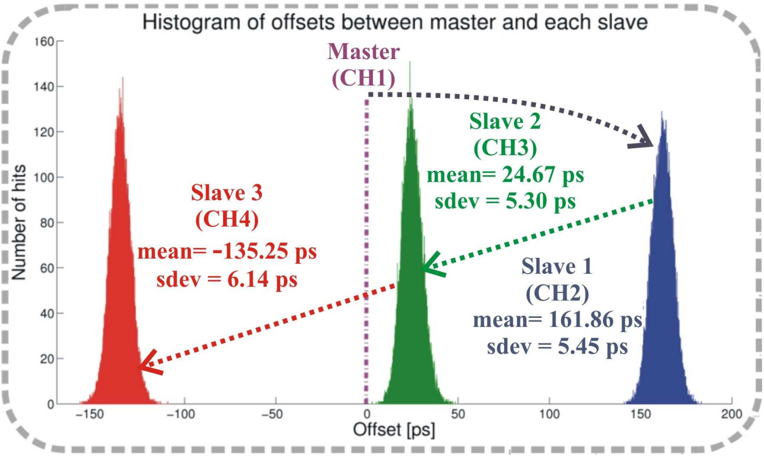
WR synchronization performance



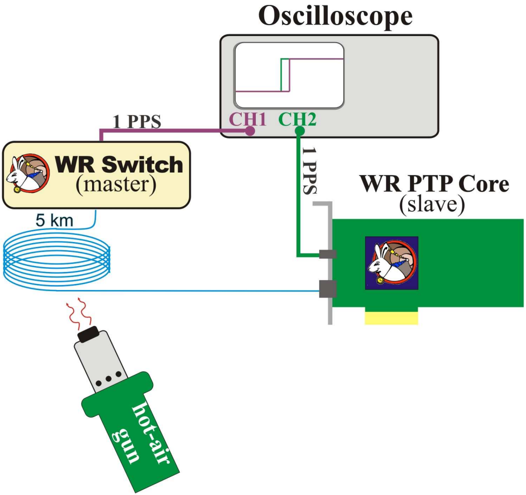
WR synchronization performance



WR synchronization performance



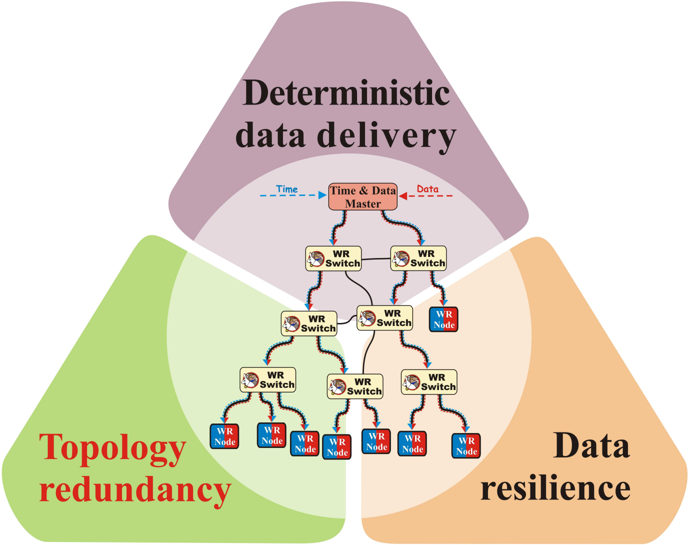
Timing demo


Oscilloscope
CH1
CH2
1 PPS
1 PPS
WR Switch
(master)
5 km
WR PTP Core
(slave)
hot-air
gun

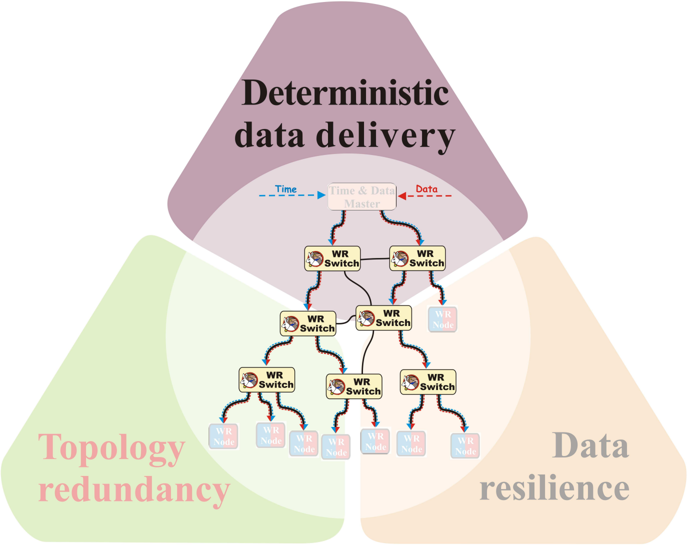
Outline

1 Introduction
2 CERN Control & Timing
3 WR Network
4 Time Distribution
- Timing demo
5 Data Distribution
- Redundancy demo
6 WR @ CERN
7 Summary

Data Distribution in a White Rabbit Network

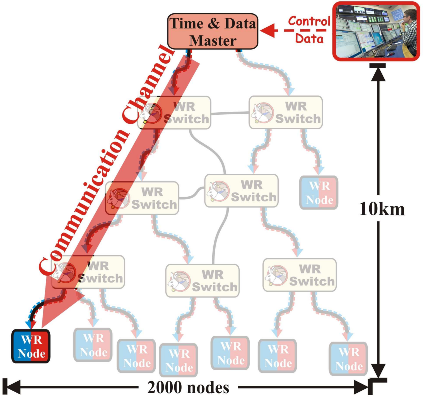


Determinism and Latency (Switch)



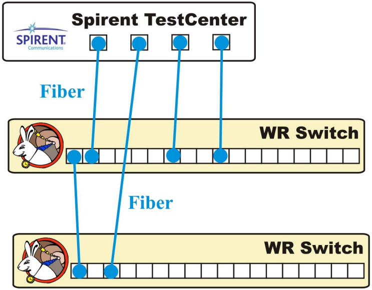
High Priority

- Types of data distinguished by 802.1Q tag:
  - **High Priority** (strict priority)
  - Standard Data (Best Effort)
- **High Priority** characteristics:
  - Broadcast/Multicast
  - Low-latency
  - Deterministic
  - Uni-directional
  - Re-transmission excluded
- Failure of **High Priority**:
  - Medium imperfection
  - Network element failure
  - Exceeded latency

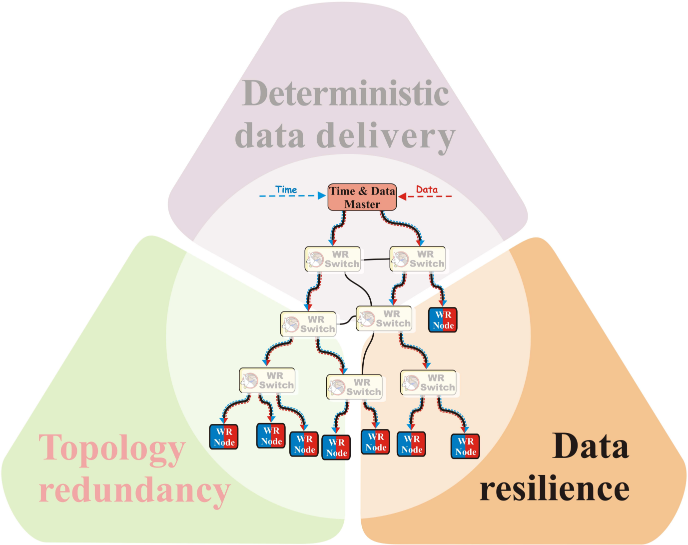


Determinism and Latency

- Deterministic Latency of High Priority
  - By design: $<$ 10us
    (single source of High Priority)
  - All size of frames
  - All rates
  - Regardless of Best Effort traffic
- Preliminary tests: $\approx$ 3us



Data Resilience (Node)



Data Redundancy

- **Forward Error Correction (FEC)** – transparent layer:
  - One message encoded into N Ethernet frames
  - Recovery of message from any M (M<N) frames

Data Redundancy

- **Forward Error Correction (FEC)** – transparent layer:
  - One message encoded into N Ethernet frames
  - Recovery of message from any M (M<N) frames
- FEC can prevent data loss due to:

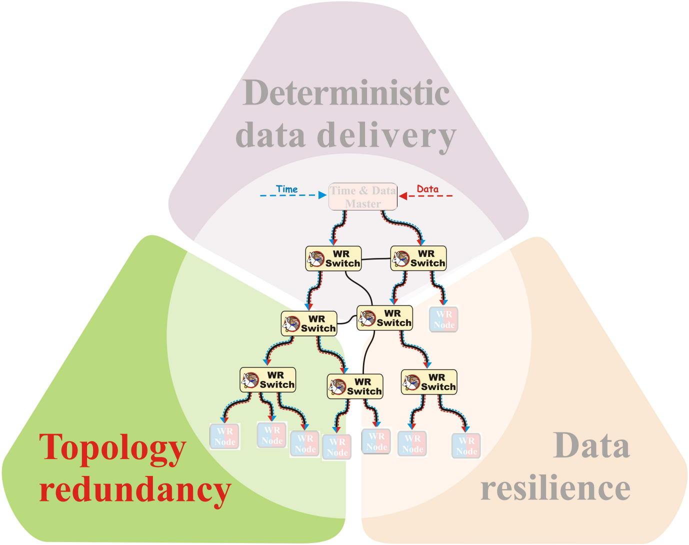
Data Redundancy

- **Forward Error Correction (FEC)** – transparent layer:
  - One message encoded into N Ethernet frames
  - Recovery of message from any M (M$<$N) frames
- FEC can prevent data loss due to:
  - **bit error**

Data Redundancy

- **Forward Error Correction (FEC)** – transparent layer:
  - One message encoded into N Ethernet frames
  - Recovery of message from any M (M<N) frames
- FEC can prevent data loss due to:
  - **bit error**
  - **network reconfiguration**

Topology Redundancy (Switch)

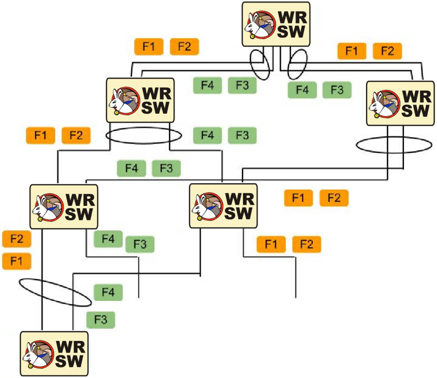


Topology Redundancy (Switch)

- Ideas:
  - Enhanced Link Aggregation Control Protocol (eLACP)
  - WR Rapid Spanning Tree Protocol (WR RSTP)
  - WR Shortest Path Bridging (WR SPB)
- Seamless redundancy = FEC + WR RSTP/SPB/eLACP
- Redundant data received in end stations
- Take advantage of broadcast/multicast characteristic of Control Data traffic (within VLAN)

Topology Redundancy: eLACP (short explanation)

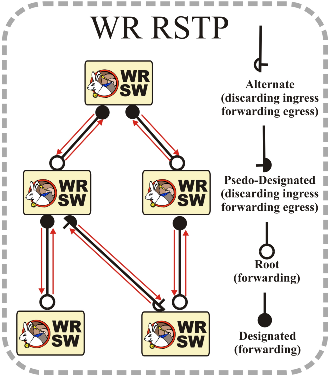
Control Message encoded into 4 Ethernet Frames (F1,F2,F3,F4). Reception of any two enables to recover Control Message (*Cesar Prados, GSI).*



Courtesy of Cesar Prados

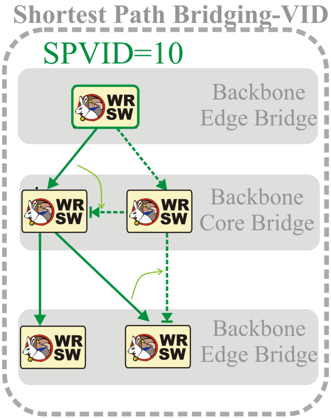
Topology Redundancy: WR RSTP

- Speed up RSTP – max 2 frames lost on re-configuration
- H/W switch-over to the backup link
- RSTP's a priori information (alternate/backup) used
- Limited number of allowed topologies
- Drop only on reception – within VLAN



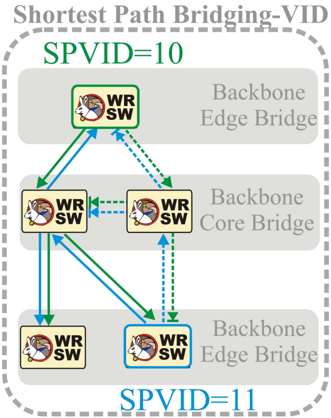
Topology Redundancy: WR SPB

- Shortest Path Bridging – VID (SPBV)
- Backup ports blocking on reception
- Single port forwarding from source
- H/W switch-over to path equally or more distant to the root



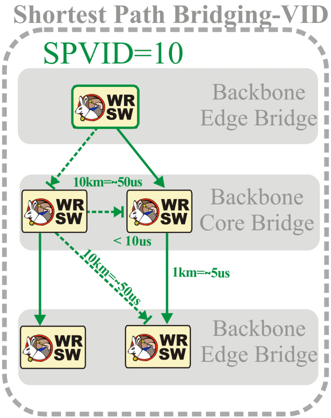
Topology Redundancy: WR SPB

- Shortest Path Bridging – VID (SPBV)
- Backup ports blocking on reception
- Single port forwarding from source
- H/W switch-over to path equally or more distant to the root
- Not fully congruent



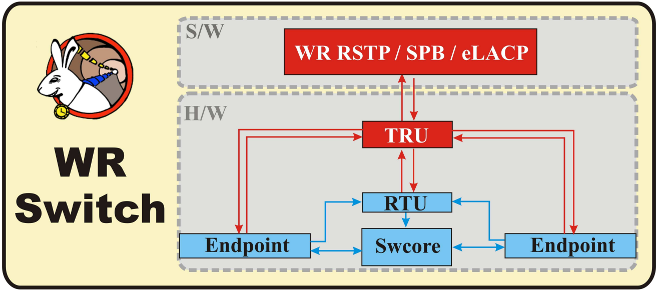
Topology Redundancy: WR SPB

- Shortest Path Bridging – VID (SPBV)
- Backup ports blocking on reception
- Single port forwarding from source
- H/W switch-over to path equally or more distant to the root
- Not fully congruent
- New link metrics: link delay

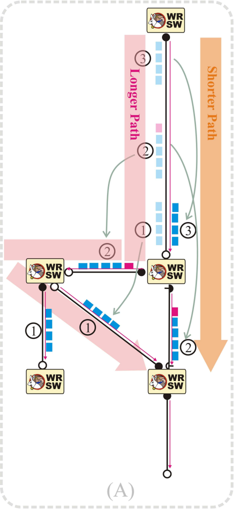


Topology Resolution Unit (TRU)

- Configurable module to support various software protocols
- Accepts active and backup port masks (ingress and egress)
- Monitors and controls ports state
- Takes actions on HW-filtered frames and link-down
- Triggers hardware generation of frames



Topology Resolution Unit (TRU)



- Marker-based hardware-switch-over
- Hardware-generated priority-based PAUSE
- Hardware-generated BPDUs
- Hardware-detection of BPDUs to open blocking (pre-configured) port

Other features/ideas

- Semi-automatic reconfiguration

Other features/ideas

- Semi-automatic reconfiguration
- Time-triggered reconfiguration

Other features/ideas

- Semi-automatic reconfiguration
- Time-triggered reconfiguration
- Time-aware shaper

Other features/ideas

- Semi-automatic reconfiguration
- Time-triggered reconfiguration
- Time-aware shaper
- Drop non-High Priority frames when High Priority arrives

White Rabbit and IEEE 802

- We want to be standard-compatible!

White Rabbit and IEEE 802

- We want to be standard-compatible!
- Ideas in line with Time Sensitive Networks

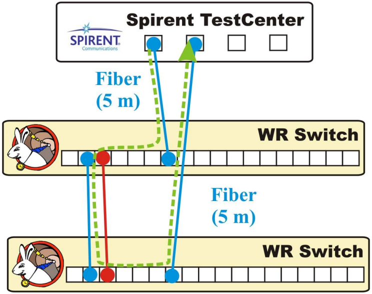
White Rabbit and IEEE 802

- We want to be standard-compatible!
- Ideas in line with Time Sensitive Networks
- Great potential for collaboration between CERN and IEEE

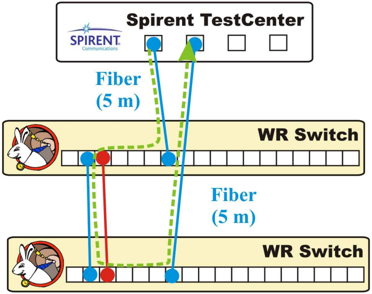
White Rabbit and IEEE 802

- We want to be standard-compatible!
- Ideas in line with Time Sensitive Networks
- Great potential for collaboration between CERN and IEEE
- Perfect platform for prototyping

Topology reconfiguration performance



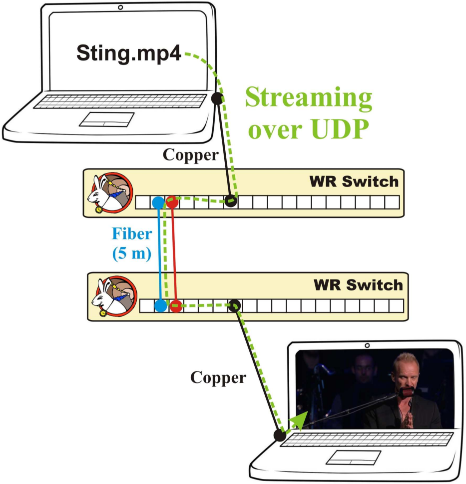
Topology reconfiguration performance



Frame Loss and Latencies

Frame Size (bytes)	Load (%)	Tx Frames	Rx Frames	Frame Loss	Max Latency (uSec)
288	10	1,217,533	1,217,533	0	5.84
288	30	3,652,598	3,652,597	1	5.84
288	50	6,087,663	6,087,663	0	5.84
288	70	8,522,728	8,522,727	1	5.84
288	90	10,957,793	10,957,792	1	6.12

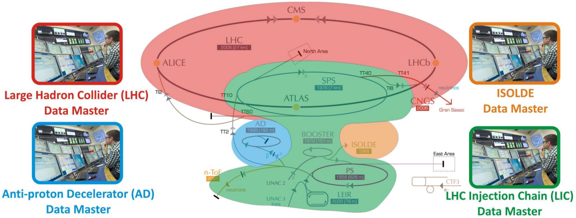
Redundancy demo



Outline

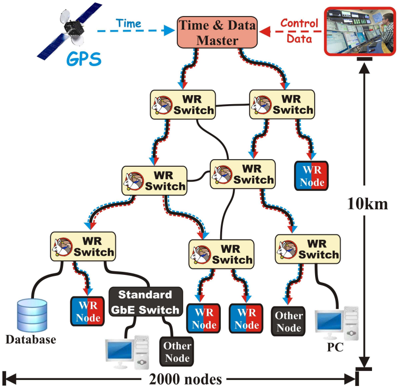
1 Introduction
2 CERN Control & Timing
3 WR Network
4 Time Distribution
   - Timing demo
5 Data Distribution
   - Redundancy demo
6 WR @ CERN
7 Summary

WR-based Control and Timing System (concept)

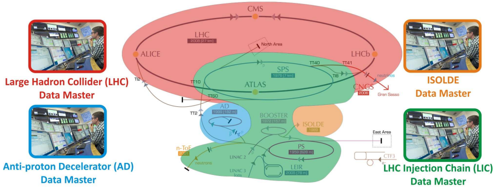


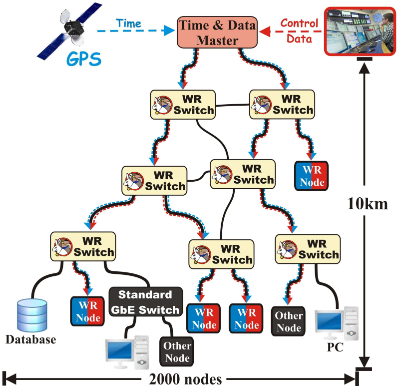
- 4 accelerator networks
- Separate **Data Master (DM)** for each network
- LIC Data Master communicates with other DMs and control devices in their networks
- Broadcast/multicast of **Control Messages**

WR-based Control and Timing System (concept)

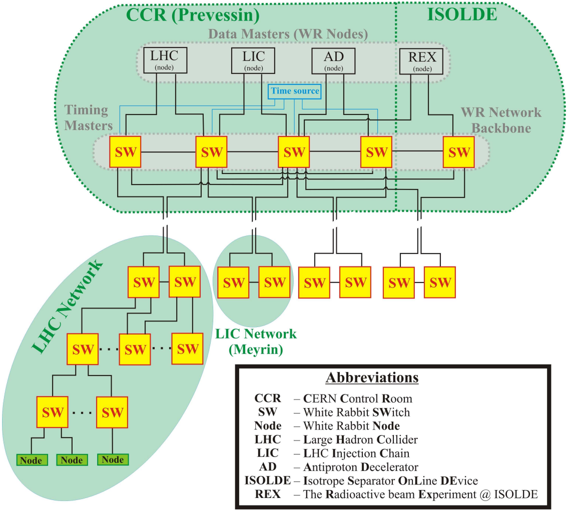


WR-based Control and Timing System (concept)





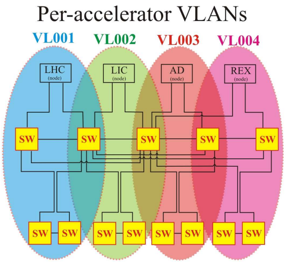
Accelerator Networks

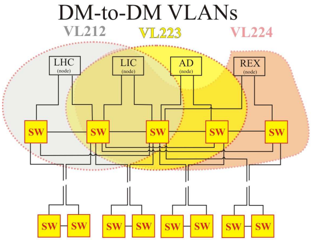


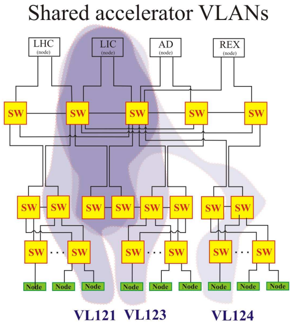
Abbreviations

CCR	– **C**ERN **C**ontrol **R**oom
SW	– White Rabbit **SW**itch
Node	– White Rabbit **Node**
LHC	– **L**arge **H**adron **C**ollider
LIC	– **L**HC **I**njection **C**hain
AD	– **A**ntiproton **D**ecelerator
ISOLDE	– **I**sotrope **S**eparator **O**n**L**ine **DE**vice
REX	– The **R**adioactive beam **EX**periment @ ISOLDE

Traffic distribution: VLANs + multicast







Abbreviations

SW	– White Rabbit SWitch	AD	– Antiproton Decelerator
LHC	– Large Hadron Collider	ISOLDE	– Isotrope Separator OnLine DEvice
LIC	– LHC Injection Chain	REX	– The Radioactive beam Experiment @ ISOLDE
DM	– Data Master		

Outline

1 Introduction

2 CERN Control & Timing

3 WR Network

4 Time Distribution
- Timing demo

5 Data Distribution
- Redundancy demo

6 WR @ CERN

7 Summary

White Rabbit Family

Successful international collaboration of institutes, universities and companies

WR Users:
http://www.ohwr.org/projects/white-rabbit/wiki/WRUsers

White Rabbit Family

Successful international collaboration of institutes, universities and companies

WR Users:
http://www.ohwr.org/projects/white-rabbit/wiki/WRUsers

Pushing frontiers

- Scientific, open (H/W & S/W), with companies

Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected

Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected
- A versatile solution for general control and data acquisition

Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected
- A versatile solution for general control and data acquisition
- Fulfilling all our needs in synchronization and determinism

Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected
- A versatile solution for general control and data acquisition
- Fulfilling all our needs in synchronization and determinism
- Standard-compatible and standard-extending

Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected
- A versatile solution for general control and data acquisition
- Fulfilling all our needs in synchronization and determinism
- Standard-compatible and standard-extending
- Active participation in IEEE1588 revision process

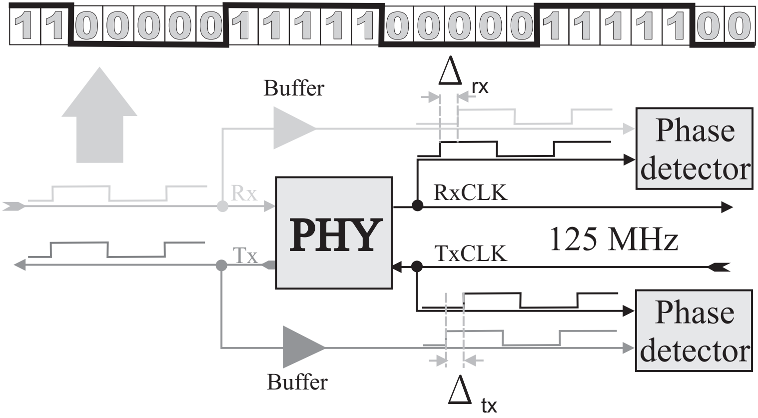
Pushing frontiers

- Scientific, open (H/W & S/W), with companies
- More applications than ever expected
- A versatile solution for general control and data acquisition
- Fulfilling all our needs in synchronization and determinism
- Standard-compatible and standard-extending
- Active participation in IEEE1588 revision process
- Eager to collaborate with IEEE802

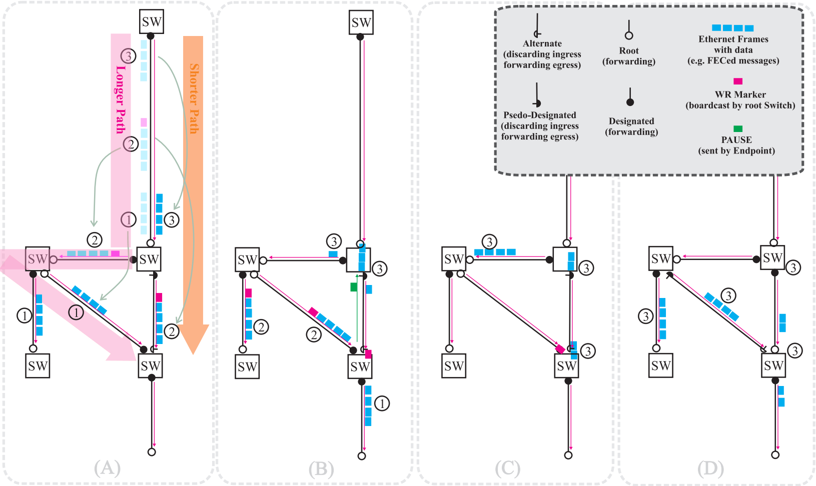
Thank you

More information:
http://www.ohwr.org/projects/white-rabbit/wiki

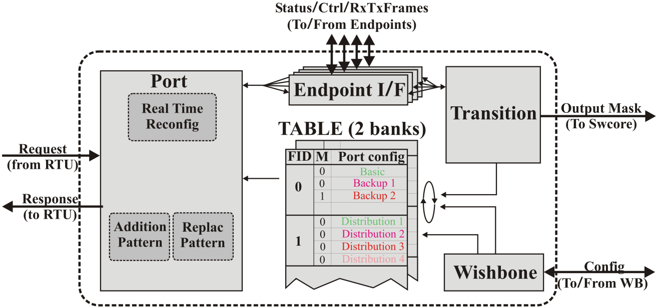
Fixed Delays Measurement



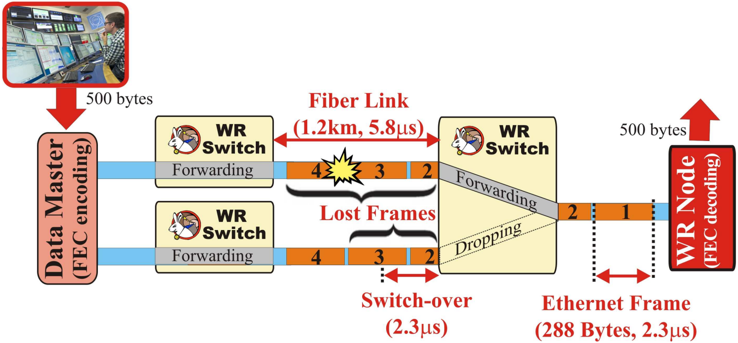
WR RSTP: adding new network element



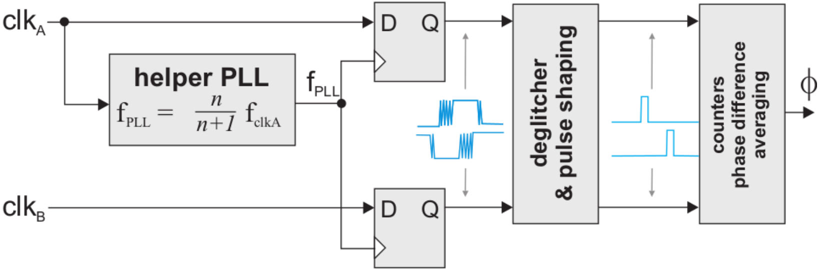
Topology Resolution Unit (TRU)



WR RSTP + FEC

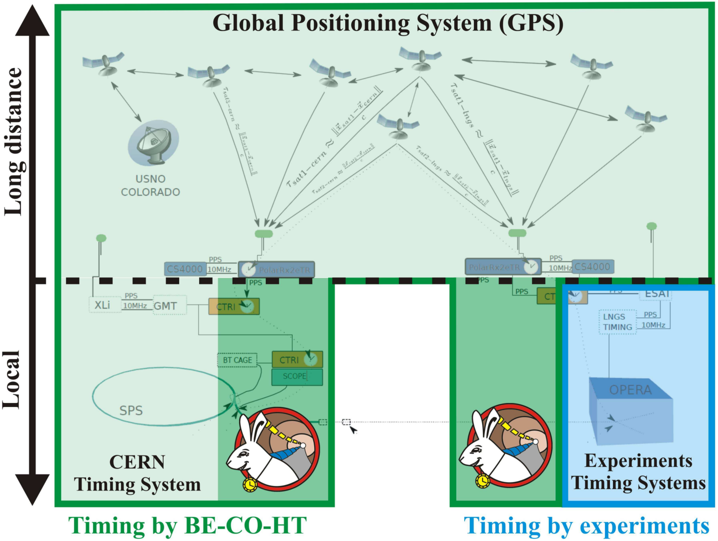

500 bytes
Fiber Link
(1.2km, 5.8µs)
WR Switch
Forwarding
WR Switch
Forwarding
WR Switch
Forwarding
Dropping
Data Master
(FEC encoding)
4
3
2
4
3
2
2
1
Lost Frames
500 bytes
WR Node
(FEC decoding)
Switch-over
(2.3µs)
Ethernet Frame
(288 Bytes, 2.3µs)

Digital Dual Mixer Time Domain (DMTD) phase detector



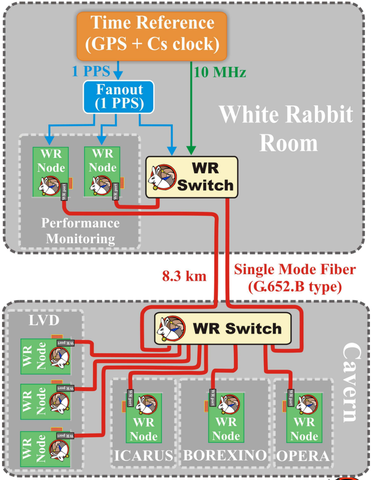
- Fully digital, so fully linear
- Can handle multiple channels without need for extra hardware

New time transfer with WR for CNGS

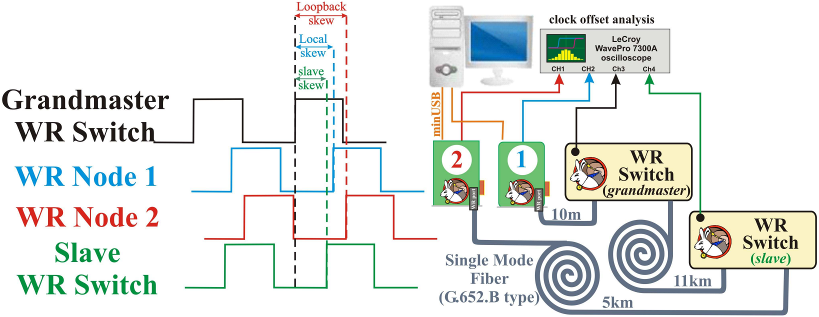


WR installation for CNGS

- Grandmaster WR Switch
- 8 km of fiber between switches
- Boundary Clock WR Switch
- WR Node – includes Time-to-Digital Converter (TDC):
  - 55 ps precision (std. dev)
  - 300 ps accuracy
- Performance monitoring

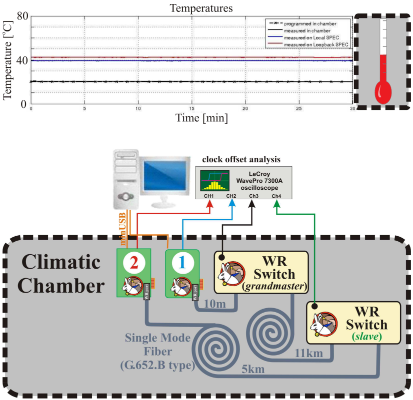


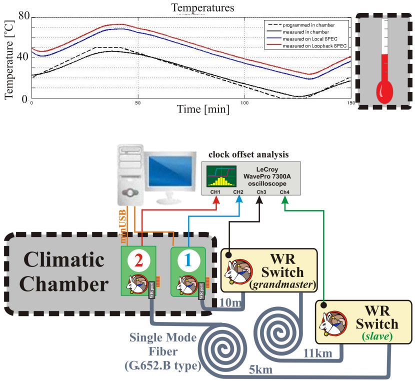
Temperature tests setup (1)



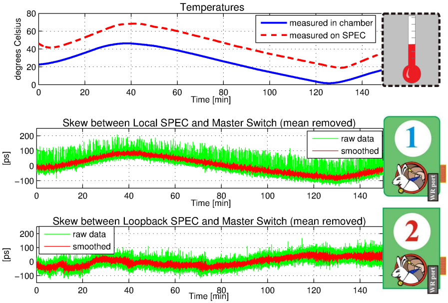
- Measurement of WR Timebase (clock)
- Skew measurement with oscilloscope

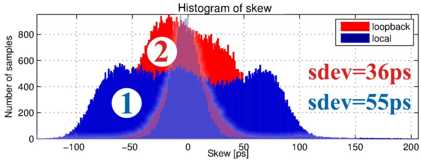
Temperature tests setup (2)

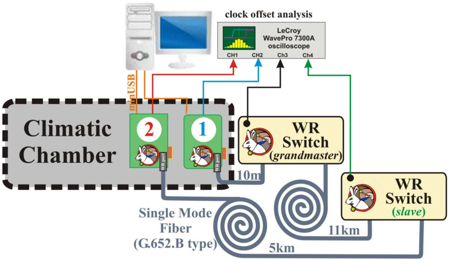




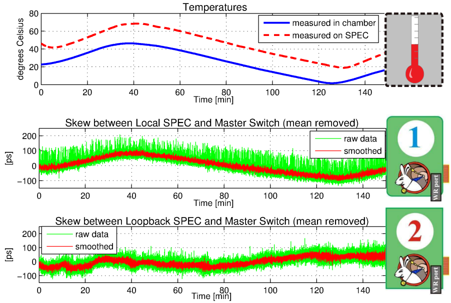
Temperature tests results (1)

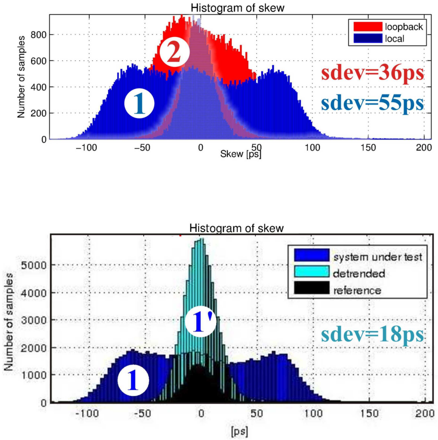






Temperature tests results (2)





The change of time offset
due to temperature changes

$\approx$ **4 ps per 1°C**