Robustness in White Rabbit

Maciej Lipinski

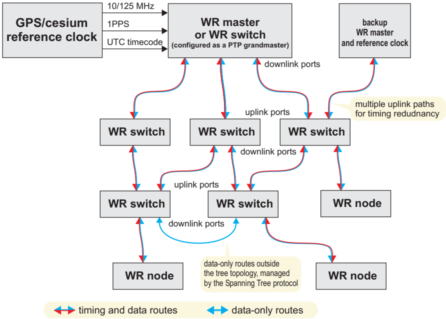
BE-CO Hardware and Timing section
CERN, Geneva,
Switzerland

PERG, Institute of Electronic Systems
Warsaw University of Technology, Warsaw,
Poland

May 27, 2011

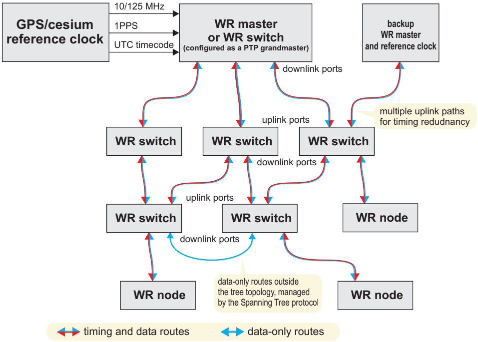
Definition

A White Rabbit Network (WRN) is considered robust only if all the WR nodes connected to the network always receive data on time and are always synchronized with the required accuracy. The amount of lost frames in a given period of time never exceeds the upper bound.



Naming Conventions

- Granularity Window (GW).
- Information distributed over WRN:
  - Control Data - Control Messages (CM),
  - Clock - WR PTP + SyncE,
  - Standard Data - all the other traffic,
- Class of Service and Quality of Service (CoS and QoS),
- High Priority traffic (HP),
- Standard Priority traffic (SP).
- Forward Error Correction (FEC).



Areas of Consideration

- Determinism
- Clock Resilience
- Data Resilience
- Monitoring and Diagnostics

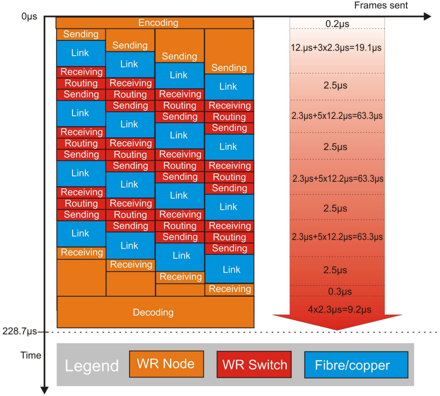
Requirements

Requirement	**Value(s)**	
	GSI	CERN
Granuality Window	100μs	1000μs
Maximum Link Length	2km	10km
Control Message Size	200-500 bytes	1200 - 5000 bytes
Synchronization accuracy	probably 8ns	most nodes 1μs few nodes 2ns
Control Message loss rate	1 per year (?)	1 per year (?)

Delivery Delay Estimation

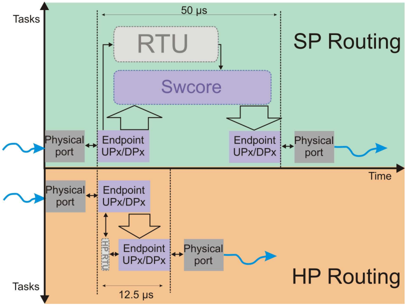
- Control Message of 500 Bytes is encoded with FEC into 4 Ethernet frames of 288 Bytes.
- Store-and-Forward implemented in SWCore is not sufficient for GSI's GW (100*us*).

CM size	CM Delivery Delay	
	GSI	CERN
200 bytes	92.2μs	132.2μs
500 bytes	228.7μs	268.3μs
1500 bytes	272.2μs	312.2μs
5000 bytes	349.3μs	389.3μs



Cut-through HP Bypass

- All the broadcast traffic with priority 7 is cut-through forwarded using HP Bypass.
- Ideas concerning HP traffic collisions :
  - Single source of HP Traffic.
  - Priority of HP Traffic from Data Master (DM), drop non-DM on collision.

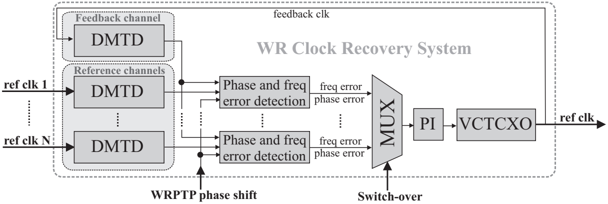


CM size	CM Delivery Delay	
	GSI	CERN
200 bytes	63.2μs	103.2μs
500 bytes	76.3μs	116.3μs
1500 bytes	106.4μs	146.4μs
5000 bytes	175.8μs	215.8μs

Synchronization Stability

What might cause synchronization instability?

- Changing conditions (e.g. temperature) – solved by WRPTP.
- Failure of network elements – solved by topology redundancy and WRPTP,
- Switch-over (change of clock source-port). Two dependencies:
  - Syntonization – SyncE - PLLs designed to accommodate many clock sources,
  - Synchronization – specially modified BMC in WRPTP.



Probability of WRN failure

Requirement name	Value(s)	
	GSI	CERN
max Failure rate ($\lambda_{WRN_{max}}$)	$3.170979198 * 10^{-12}$	$3.170979198 * 10^{-11}$

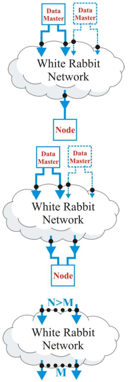
$$P_{WRN_f} = P_{congestion} + P_{f\_FEC} + P_{f\_Network} \quad (1)$$

- $P_{congestion}$ - Control Message lost (dropped) due to congestion.
- P_{f_FEC} - FEC fails to recover Control Message.
- $P_{f\_Network}$ - single network component failure.

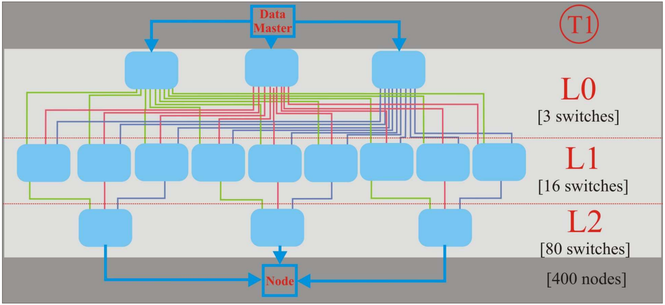
	WRS Number	**Nodes MAX Number**	$MTBF_{Switch}$ **= 20 000[h]**	
Topology			P_f	MTBF[h]
No-redundant	127	2048	$2.08 * 10^{-3}$	$5.77 * 10^{3}$
Double-redundancy	292	2048	$4.71 * 10^{-7}$	$2.55 * 10^{7}$
Triple-redundancy	495	2048	$3.06 * 10^{-11}$	$4.08 * 10^{11}$

Topology Redundancy

- Increases Clock and Data resilience by eliminating *single point of failure* (only if redundant connection to WR Node is considered).
- Enables to achieve reliability of entire network greater then reliability of its single component.
- First estimations show that double redundancy is not enough to achieve reliability of 1 CM lost per year, TO BE confirmed with more studies.
- The redundancy of the WRN is justified only if Data Master is highly reliable or redundant.



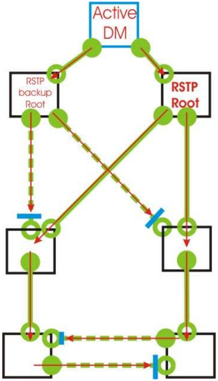
Triple-redundancy of topology



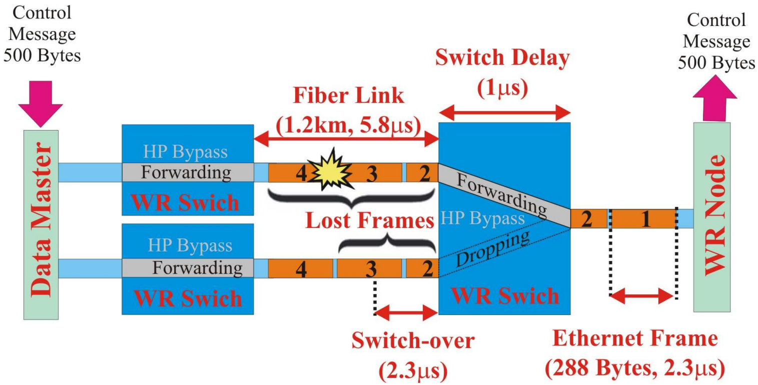
- For ≈2000 WR nodes connected to two layers of switches, 15 switches in L0, 80 in L1 and 400 in L2 are required (total 495)

Rapid Spanning Tree Protocol in WR (WR RSTP)

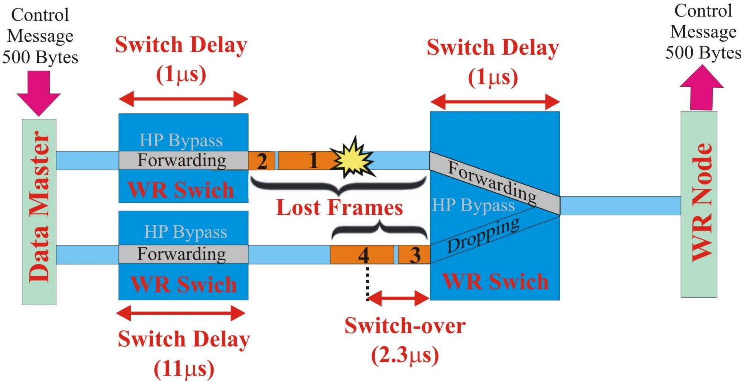
- Requirements:
  - Fast switching to alternate/backup link so that not more then 2 HP Frames are lost, e.g.: CM of 500 Bytes, FECed into 4 Ethernet frames of 288 Bytes, each transmitted $2.3us$ – switching time $< 2.3us$
  - Alternate path length : max 1 hop longer than the primary path length.
- The speed of White Rabbit RSTP is directly associated with the minimum CM size.
- Hardware support for HP traffic (only) using RSTP and restricting possible topologies.
- Challenge: WR RSTP for all the Ethernet traffic



WR RSTP – theoretical consideration



WR RSTP – real-life consideration

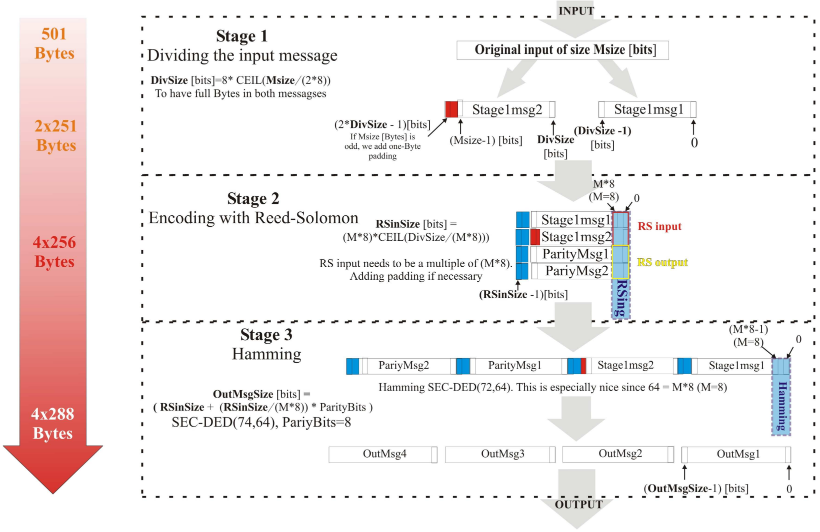


- Introducing maximum cut-through delay (13*us*) on backup ports of the switch.
- Backup link always 1 hoop longer then active.

Data Redundancy (FEC)

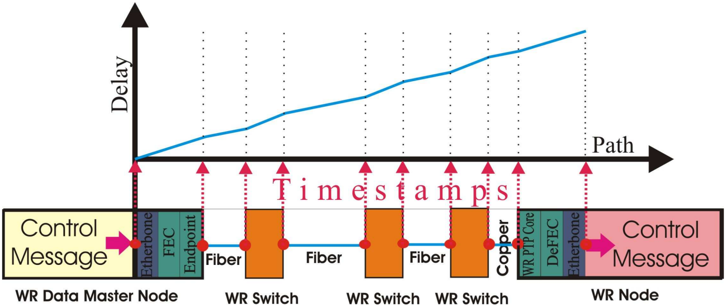
- Reed-Solomon for package-based encoding:
  4 Ethernet Frames (2 x original, 2 x parity) for input of size $<\approx 2500$. We can lose any 2 packages (out of 4).
- Hamming for bit-based encoding – Single Error Detection-Double Error Correction (SEC-DED).

Flow and Congestion Control



Monitoring and Diagnostics of WR-specific parameters

- Detection of lost HP Frames (in WR Switches) using FEC ID and CM ID (stored in the header added by FEC).
- Precise knowledge of HP traffic delays on the path DataMaster $< - >$ Node.
- Monitoring of WRPTP parameters.



Thank you

Thank you for your attention

Questions?