Robustness in White Rabbit

Maciej Lipinski, Cesar Prados

CERN & GSI

14/04/2011
Robustness

What is a robust White Rabbit Network
Naming Conventions
Areas of Consideration
Requirements

Areas of Consideration
Determinism
Clock Resilience
Data Resilience
Monitoring and Diagnostics

DISCUSSION (OHMG)
Oh, My God !!!
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).
- Types of the information distributed over WRN:
  - Control Data - Control Messages (CM),
  - Clock - timing information (PTP+SyncE),
  - Standard Data - all the other Ethernet traffic,
- Class of Service and Quality of Service (CoS and QoS),
- High Priority traffic (HP),
- Standard Priority traffic (SP).

DISCUSSION:
Most of the names introduced here are not in line with standard naming conventions and need to be changed, ideas appreciated.
Areas of Consideration

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

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granularity Window</td>
<td>GSI: 100µs</td>
</tr>
<tr>
<td></td>
<td>CERN: 1000µs</td>
</tr>
<tr>
<td>Maximum Link Length</td>
<td>GSI: 2km</td>
</tr>
<tr>
<td></td>
<td>CERN: 10km</td>
</tr>
<tr>
<td>Control Message Size</td>
<td>GSI: 200-500 bytes</td>
</tr>
<tr>
<td></td>
<td>CERN: 1200 - 5000 bytes</td>
</tr>
<tr>
<td>Synchronization accuracy</td>
<td>GSI: probably 8ns</td>
</tr>
<tr>
<td></td>
<td>CERN: most nodes 1µs</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Message loss rate</td>
<td>GSI: 1 per year (?)</td>
</tr>
<tr>
<td></td>
<td>CERN: 1 per year (?)</td>
</tr>
</tbody>
</table>

### DISCUSSION:

What is the real-life requirement for number of CM lost per year? GW of 100µs is very tight and needs solid justification.
Control Message of 500 Bytes is encoded with FEC into 4 Ethernet frames of ≈ 375 Bytes (current FEC 4x288 Bytes)

Store-and-Forward implemented in SWCore is not sufficient for GSI’s GW (100μs).

<table>
<thead>
<tr>
<th>CM size</th>
<th>CM Delivery Delay GSI</th>
<th>CM Delivery Delay CERN</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 bytes</td>
<td>92.2μs</td>
<td>132.2μs</td>
</tr>
<tr>
<td>500 bytes</td>
<td>228.7μs</td>
<td>268.3μs</td>
</tr>
<tr>
<td>1500 bytes</td>
<td>272.2μs</td>
<td>312.2μs</td>
</tr>
<tr>
<td>5000 bytes</td>
<td>349.3μs</td>
<td>389.3μs</td>
</tr>
</tbody>
</table>
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.

**DISCUSSION:**
Drop on collision, sending of HP by more Data Master and/or nodes.

<table>
<thead>
<tr>
<th>CM size</th>
<th>CM Delivery Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GSI</td>
</tr>
<tr>
<td>200 bytes</td>
<td>63.2µs</td>
</tr>
<tr>
<td>500 bytes</td>
<td>76.3µs</td>
</tr>
<tr>
<td>1500 bytes</td>
<td>106.4µs</td>
</tr>
<tr>
<td>5000 bytes</td>
<td>175.8µs</td>
</tr>
</tbody>
</table>
Clock instability due to changing conditions prevented by PTP.

Topology redundancy, logical topology of Clock distribution is aligned with logical topology of Data, active and backup ports (clock/data sources) assigned using Rapid Spanning Tree Protocol.

Clock Stability preservation during switch-over (change of clock source-port). Two dependencies:

- Syntonization – SyncE - PLLs by Tomek are designed to accommodate many clock sources, theoretically stability during switch-over is not an issue.
- Synchronization – WRPTP - Each port speaks independently WRPTP unaware which one is active, no change WRPTP-wise during switch over.

Tests are needed to prove the theory in practice.
## Probability of WRN failure

$$P_{WRN_f} = P_{congestion} + P_{f\_FEC} + P_{f\_Network}$$

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

<table>
<thead>
<tr>
<th>Topology</th>
<th>WRS Number</th>
<th>Nodes MAX Number</th>
<th>MTBF$_{Switch}$= 20 000[h]</th>
<th>$P_f$</th>
<th>MTBF[h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-redundant</td>
<td>127</td>
<td>2048</td>
<td></td>
<td>$2.08 \times 10^{-3}$</td>
<td>$5.77 \times 10^{-4}$</td>
</tr>
<tr>
<td>Double-redundancy</td>
<td>292</td>
<td>2048</td>
<td></td>
<td>$4.71 \times 10^{-7}$</td>
<td>$2.55 \times 10^{-7}$</td>
</tr>
<tr>
<td>Triple-redundancy</td>
<td>495</td>
<td>2048</td>
<td></td>
<td>$3.06 \times 10^{-11}$</td>
<td>$4.08 \times 10^{11}$</td>
</tr>
</tbody>
</table>
Topography 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.

**DISCUSSION:**
Redundant connection to WR nodes.
For \( \approx 2000 \) 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 than 2 HP Frames are lost, e.g.: CM of 500 Bytes, FECed into 4 Ethernet frames of 288 Bytes, each transmitted $2.3\mu s$ – switching time $< 2.3\mu s$
  - Alternate path length: max 1 hop longer 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.

**DISCUSSION:**

Ring topology - impossible with current approach, is it worth considering, what is the feasibility of ring WRN?
WR RSTP – theoretical consideration
WR RSTP – real-life consideration

- Introducing maximum cut-through delay (13\(\mu\)s) 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).
  - $?????$ for $>\approx 2500$


- FEC needs to know the size of incoming CM at the beginning of the encoding (needs to be provided in non-standard way, i.e. user defined register, WB address 0x2)

**DISCUSSION:**

How do we want to transfer CM $> 1500$Bytes (i.e. 5000Bytes)? Burst? non-standard size of the payload?

How do we indicate that FEC is to be used? EtherType? WB special reg?
Flow and Congestion Control

Stage 1
Dividing the input message

\[ \text{DivSize} \text{[bits]} = 8 \times \text{CEIL(Msize/(2*8))} \]

To have full Bytes in both messages

Stage 2
Encoding with Reed-Solomon

\[ \text{RSinSize} \text{[bits]} = (M*8) \times \text{CEIL(DivSize/(M*8))} \]

RS input needs to be a multiple of (M*8),
Adding padding if necessary

Stage 3
Hamming

\[ \text{OutMsgSize} \text{[bits]} = \] 
\[ (\text{RSinSize} + (\text{RSinSize}/(M*8)) \times \text{ParityBits}) \]

SEC-DED(74,64), ParityBits=8

Hamming SEC-DED(72,64). This is especially nice since 64 = M*8 (M=8)
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.
Summary of issues to be discussed

- **Naming**: Granularity Window, High Priority, Standard Priority, Control Messages.

- **Requirements**:
  - What is the real-life requirement for number of CM lost per year?
  - GW of 100µs is very tight and needs solid justification.

- **Cut-through Bypass**: Drop on collision, sending of HP traffic by more Data Master and/or nodes.

- **Topology Redundancy**: Redundant connection to WR nodes.

- **RSTP**: Ring topology - impossible with current approach, is it worth considering, what is the feasibility of ring WRN?

- **Data Redundancy (FEC)**:
  - how do we want to transfer CM > 1500 Bytes (i.e. 5000 Bytes)? Burst? non standard size payload?
  - How do we indicate that FEC is to be used? EtherType? Special WB register (addr=0x2)?

- **Flow and Congestion Control**:

- **Monitoring and Diagnostics of WR-specific parameters**: