1000BASE-T/10GBASE-T Time-To-Link and Some Implications for 2.5/5GBASE-T

IEEE P802.3bz 2.5G/5G BASE-T Task Force
Architecture ad hoc

Pete Cibula, Intel

May 5th, 2015

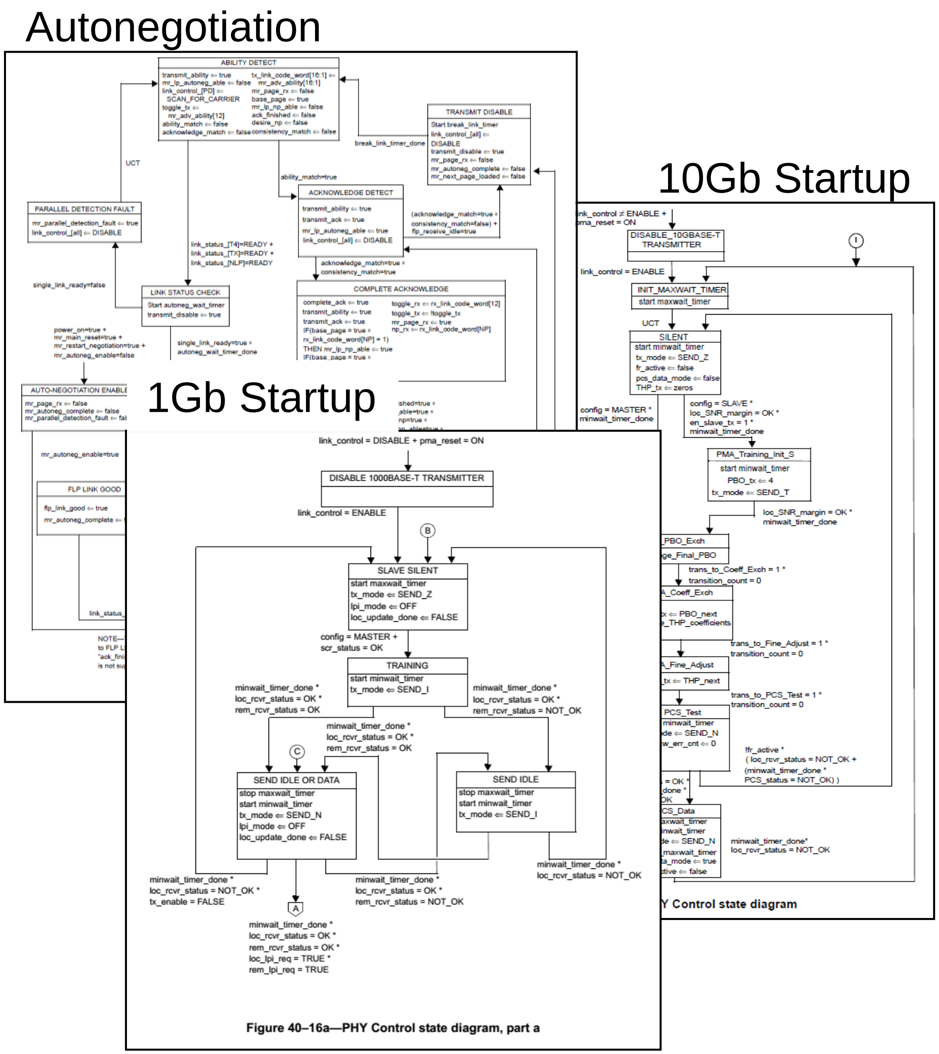
Discussion Outline

- What is time-to-link, why is it important, and how can it be characterized?
- A few representative measurements
- Factors that influence time-to-link
- Observations and summary

What is Time-To-Link (TTL)?

- Time-To-Link (TTL): A system performance metric that characterizes and measures PHY behavior through autonegotiation and 1G/10G BASE-T startup sequences
  - Autonegotiation in 802.3 Clause 28, “Physical Layer link signaling for Auto-Negotiation on twisted pair”
  - 1Gb in 802.3 Clause 40, Subclause 40.4.2.4, “PHY Control function”
  - 10Gb in 802.3 Clause 55, Subclause 55.4.2.5.14, “Startup sequence”

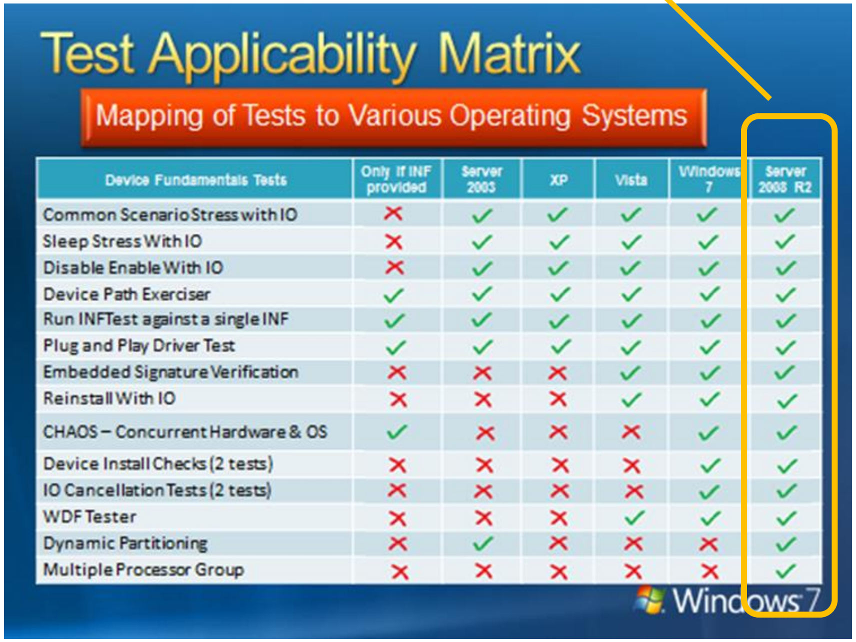
- One of two primary performance measures (along with BER) used to characterize BASE-T physical layer link interoperability



Why is it Important?

- Server networking drivers must meet 3rd-party certifications
- Example - Windows Hardware Quality Labs (WHQL) testing & certification "devfund"
  - A series of "device fundamentals" tests to evaluate the compatibility, reliability, performance, security and availability of a device in Windows OS
  - Includes many automated driver stress tests that execute multiple device resets
  - Long link times appear as a "failure" to these tests, which expect a link in 3s-4s based on 10Mb/100Mb/1Gb PHY performance

Server device fundamentals requirements



Device Fundamentals Tests	Only If INF provided	Server 2003	XP	Vista	Windows 7	Server 2008 R2
Common Scenario Stress with IO	✗	✓	✓	✓	✓	✓
Sleep Stress With IO	✗	✓	✓	✓	✓	✓
Disable Enable With IO	✗	✓	✓	✓	✓	✓
Device Path Exerciser	✓	✓	✓	✓	✓	✓
Run INFTest against a single INF	✓	✓	✓	✓	✓	✓
Plug and Play Driver Test	✓	✓	✓	✓	✓	✓
Embedded Signature Verification	✗	✗	✗	✓	✓	✓
Reinstall With IO	✗	✗	✗	✓	✓	✓
CHAOS – Concurrent Hardware & OS	✓	✗	✗	✗	✓	✓
Device Install Checks (2 tests)	✗	✗	✗	✗	✓	✓
IO Cancellation Tests (2 tests)	✗	✗	✗	✗	✓	✓
WDF Tester	✗	✗	✗	✓	✓	✓
Dynamic Partitioning	✗	✓	✗	✗	✗	✓
Multiple Processor Group	✗	✗	✗	✗	✗	✓

Source: Device Fundamentals Overview Presentation at Ihv_devfund.pptx

- Long TTLs (>6s) can lead to device certification failures!

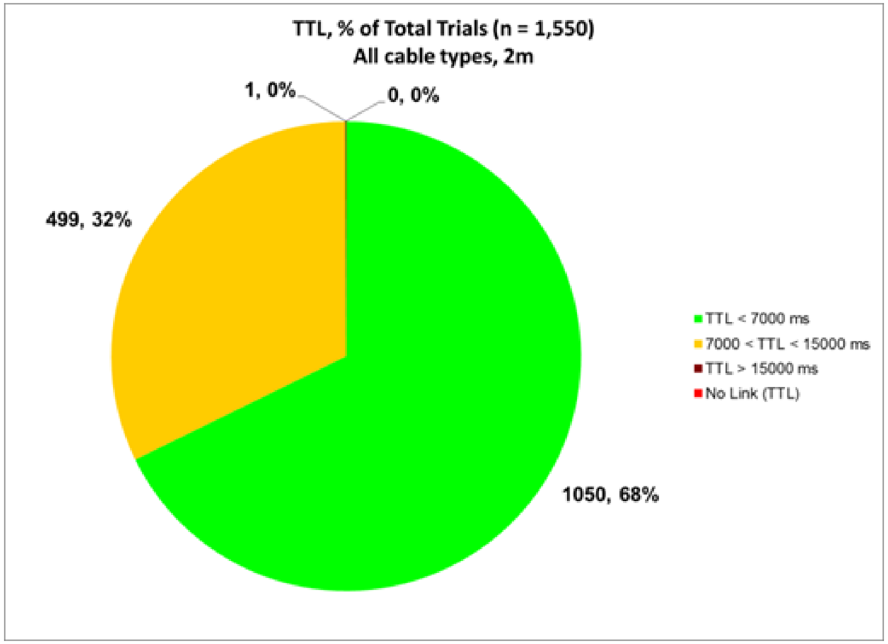
Link Interoperability Measurements

- Representative Link Interoperability metrics associated with TTL
  - Time-To-Link (Time to achieve link after link initiation event)
  - # Link Attempts (Number of attempts for each link)
  - # Link Drops (Number of link drops observed after link is established)
  - Clock Recovery (Master/Slave resolution)
  - TTL Distribution (% of links by link time)
  - Speed Downshift/Downgrade (Resolved speed if other than 10Gb/s)
- Variables that can affect TTL
  - Channel (type, configuration, length)
  - Link initiation event on either endpoint
    - Hardware reset, “soft” reset or MDIO PHY reset, autoneg restart, transmitter disable/enable, cable connect/disconnect

Characterizing Time-To-Link Behavior TTL as a Percentage of Total Trials

Total to remember: 1,550 link tests

- 1,050 out of 1,550 tests, or 68% of the total number of link tests, achieved a link state in 7s or less (green slice)
- 499 out of 1,550 tests, or 32% of the total number of link tests, achieved a link state somewhere between 7s and 15s (yellow slice)
- 1 out of 1,550 tests, or <1 % (actually 0.15%) of the total number of link tests, achieved a link state longer than 15s (actually 16.4s; can't see this in the pie chart)



Characterizing Time-To-Link Behavior Cumulative % TTL

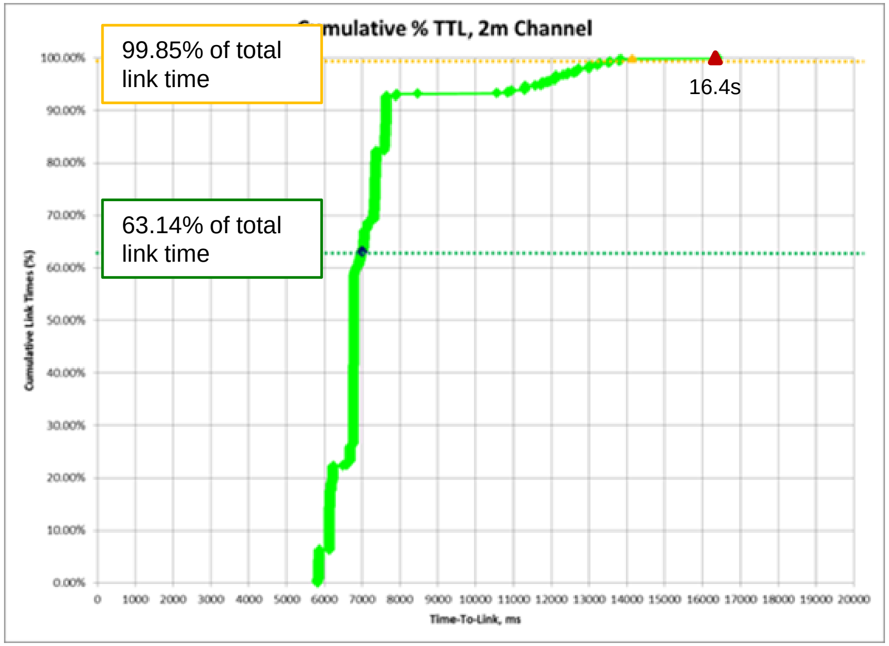
Cumulative % TTL is the distribution of measured link times as a percentage of the total measured link time

Total to remember: Total link time recorded for all 1,550 tests = 10,837,835 ms or about 3h 0min 38sec

- 1,050 tests with TTL <= 7s had a total link time of 6,843,118 ms (63.14% of the total measured link time)
- 499 tests with 7s < TTL <= 15s had a total link time of 3,978,317 ms (40.24% of the total measured link time)
- 1 test with TTL >15s had a total link time of 16,400 ms (0.15% of the total measured link time)

Expressed as a cumulative percentage

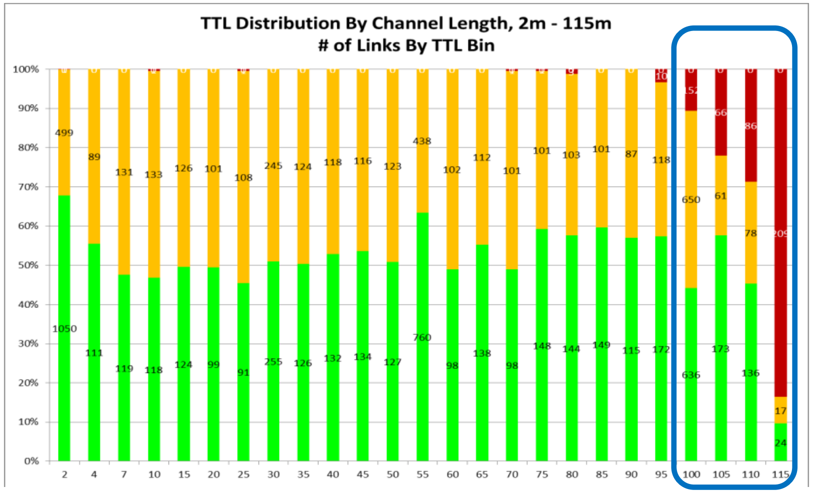
- Measured link time < 7s: 63.14%
- Measured link time <15s: 99.85%
- Measured link time <16.4s (max): 100%



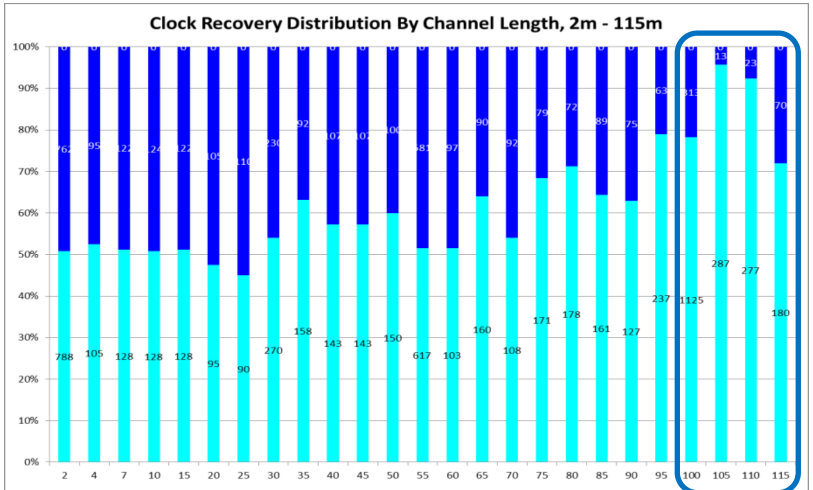
TTL bin	Total Time in TTL bin	Total Time All Tests	% Total Time	% Cum Time
<= 7s	6,843,118	10,837,835	63.14%	63.14%
7s < TTL <= 15s	3,978,317		36.71%	99.85%
> 15s	16.400		0.15%	100.00%

Example: TTL Distribution and Master/Slave Resolution by Channel Length

- Example of 10GBASE-T TTL measured from 2m to 115m channels (9,790 links)
  - Stacked plot order L-R is 2m to 115m
- TTL across 2m-100m
  - Average TTL = 7.6s
  - Average time in AN = 5s
  - Average time in training = 2.6s
- Note apparent loop timing trend towards MASTER preference with increasing channel length
- Very long TTLs (>15s) at 100m+ channels are associated with downshifts to 1Gb link speed

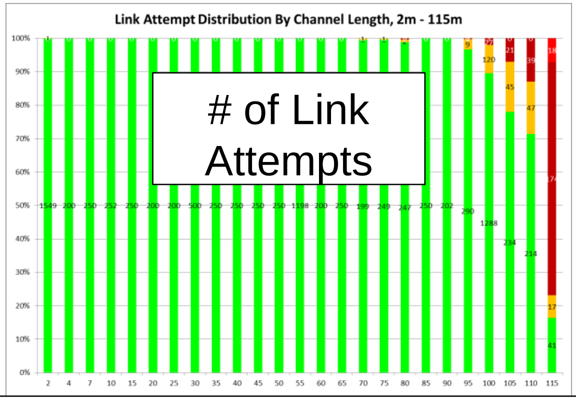


Green: TTL <= 7s; Yellow: 7s < TTL <=15s; Red: TTL >15s

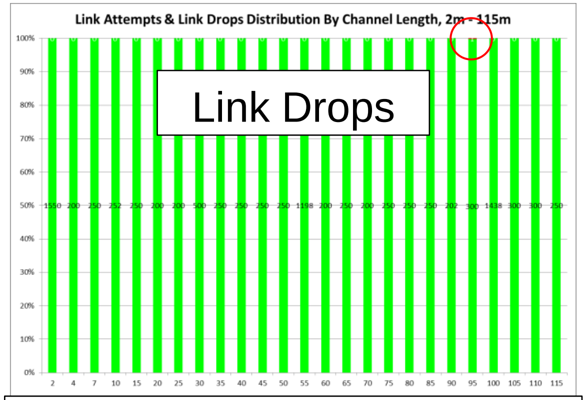


Cyan = Master; Blue = Slave

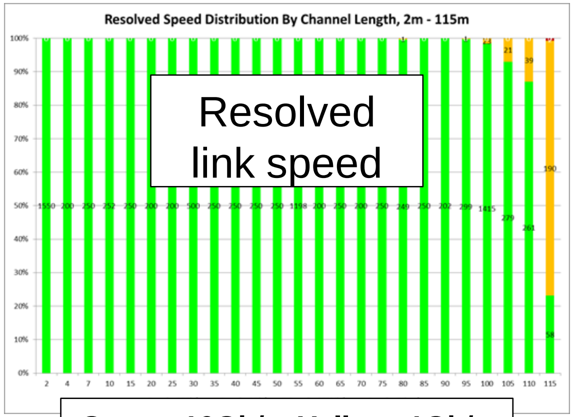
Other TTL Metrics (Same Dataset)



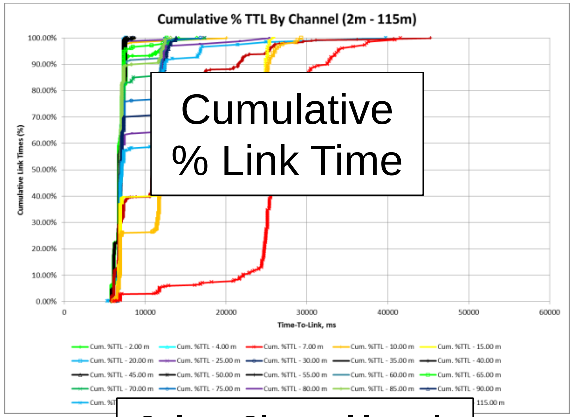
Green: 1-2; Yellow: 3-4; Crimson: 5-6; Red: >6



Green: # no link drop; Red: # link drop



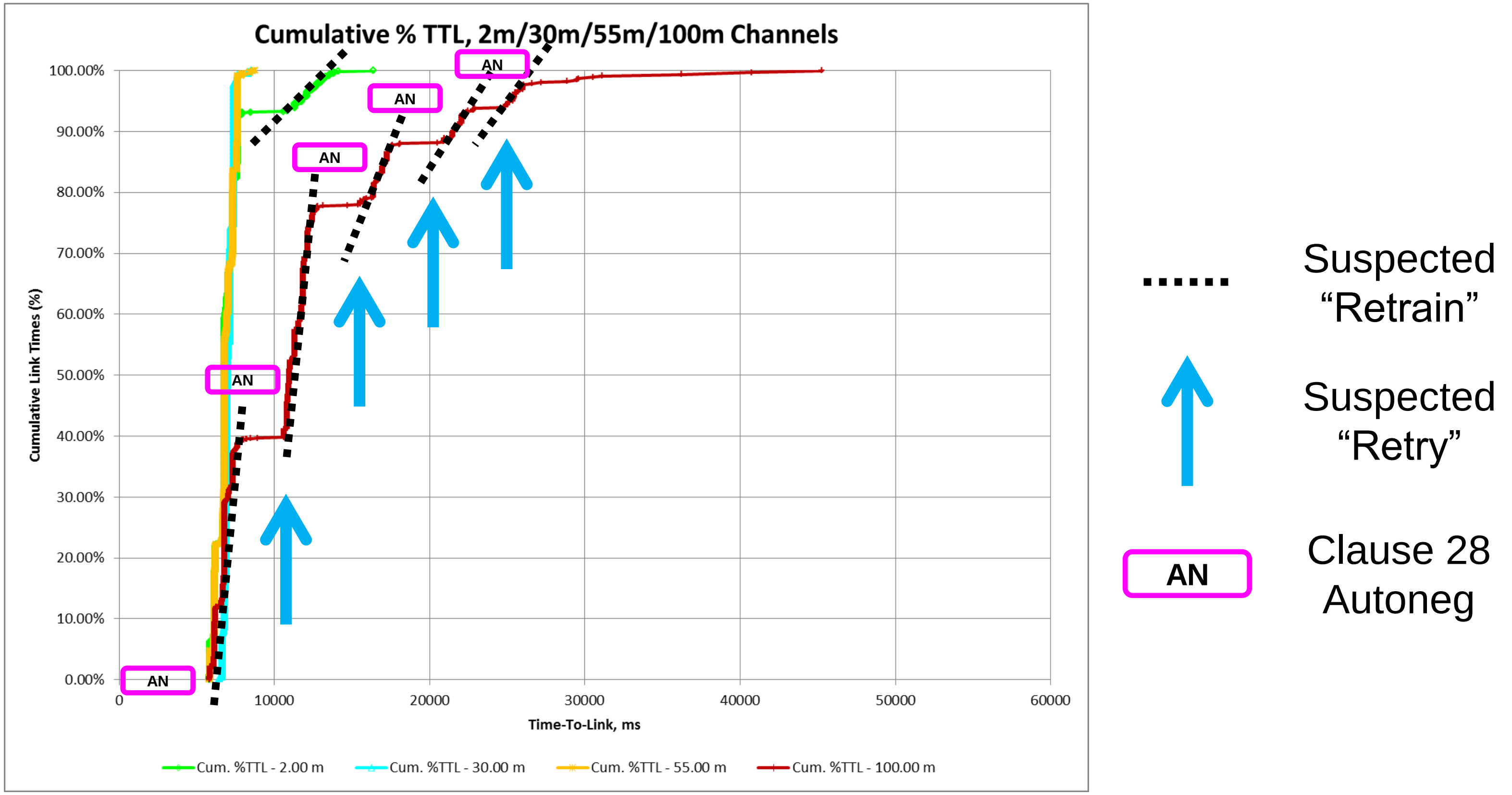
Green: 10Gb/s; Yellow: 1Gb/s



Color: Channel length

Summary: Link interoperability measurements can clearly show differences in PHY autonegotiation and link state behavior as a function of channel characteristics

Time-To-Link Levers?



- TTL is a combination of both autonegotiation and 1G/10G startup behavior
  - Three sources of variability: Autonegotiation, "Retrain" (variability through 55.4.6.1) and "Retry" (return to 28.3.4)
  - Longest TTLs typically driven by multiple passes through the Clause 28 Arbitration state diagram after failed training attempts

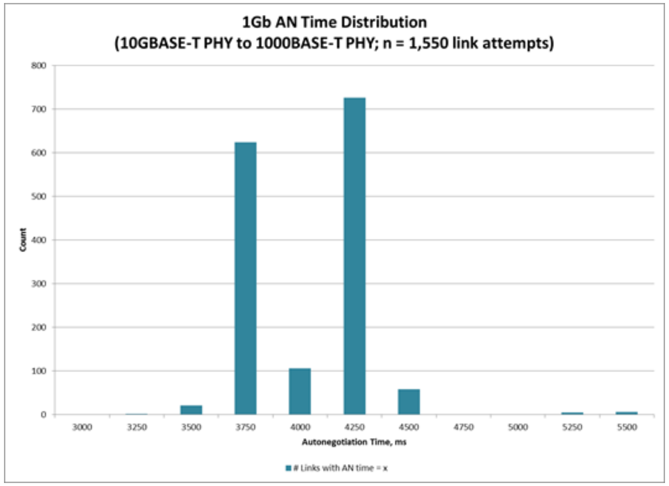
AN & Training Times: 1000BASE-T

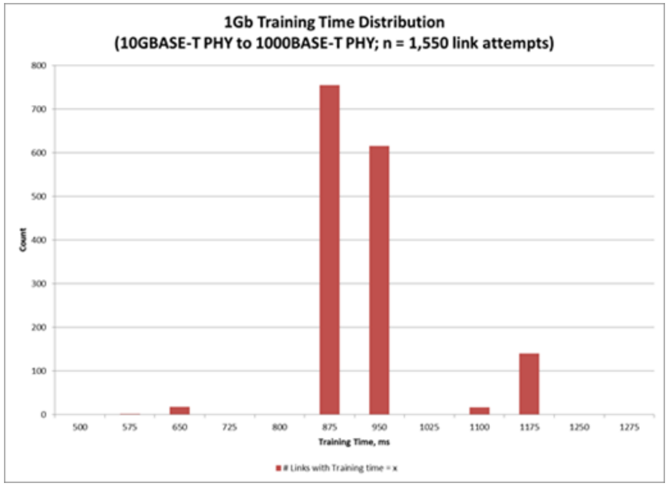
Measured autonegotiation and training times from 1,550 1Gb links.

- 10GBASE-T device to 1000BASE-T link partner
- Both endpoints set to “full auto” (autonegotiate speed, mode, and loop timing)

Results

- Autonegotiation
  - Average = 3.89s
  - Range = 3.25s to 5.50s
- Training
  - Average = 0.91s
  - Range = 0.575s to 1.175s





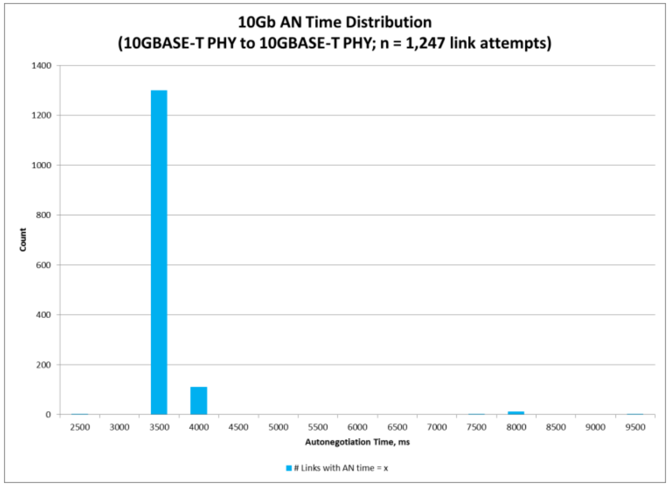
AN & Training Times: 10GBASE-T

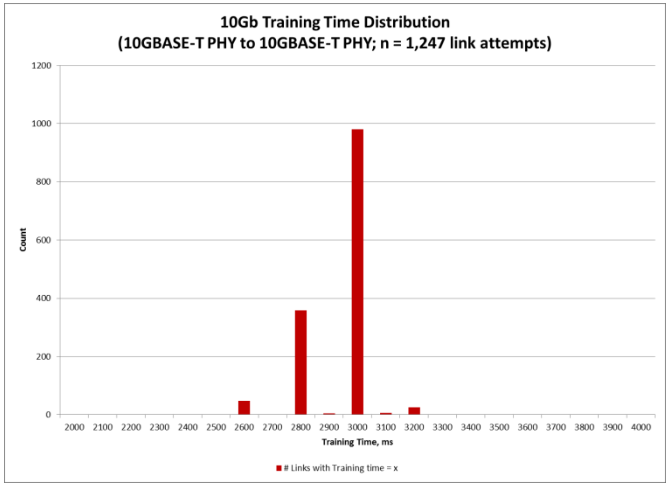
Measured autonegotiation and training times from 1,550 1Gb links.

- 10GBASE-T device to 10GBASE-T link partner
- Both endpoints set to “full auto” (autonegotiate speed, mode, and loop timing)

Results

- Autonegotiation
  - Average = 3.35s
  - Range = 2.46s to 9.24s
- Training
  - Average = 2.98s
  - Range = 2.22s to 3.66s

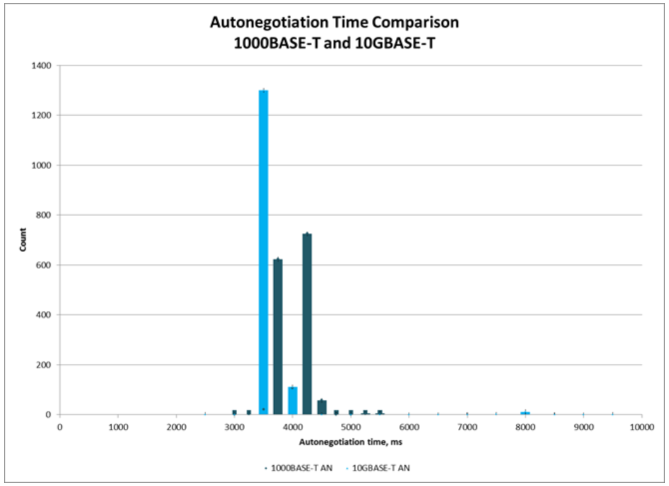




AN Comparison - 1Gb/10Gb

Current autonegotiation times for 1000BASE-T and 10GBASE-T are comparable.

From an end-user perspective, it is highly desirable that 2.5G/5GBASE-T autonegotiation times align with these technologies, and that total time-to-link be minimized.



Technology	Representative Average AN (ms)	Representative Average Training (ms)	Representative Average TTL (s)
1000BASE-T*	3889	909	4.80
2.5GBASE-T	??	??	??
5GBASE-T	??	??	??
10GBASE-T	3346	2898	6.24

**10GBASE-T to 1000BASE-T link partner. 1000BASE-T to 1000BASE-T is slightly faster.*

Observations from 10GBASE-T

- Channel topologies significantly affect the channel solutions realized by PHY DSP systems
  - "Peaky" impairments (return loss, crosstalk) appear to be a factor in link-trial-to-link-trial variability in the system solution
  - Transition region between RL/crosstalk-driven to IL-driven solutions
  - Channel lengths near 10GBASE-T PBO transitions
- PHY-specific responses to channel characteristics drive variability in autonegotiation and training time
  - Loop timing/clock recovery resolution
  - Time spent in 10GBASE-T startup states
- May have implications for both system performance and end-user experience
  - Potential to affect product time-to-market and customer ease-of-use

Summary

- Time-to-link from the end-user perspective
  - User time-to-link experience with the installed base of Cat5e/Cat6 cabling and 1000BASE-T is between 3s & 4s
  - User time-to-link experience with 10GBASE-T is ~7s (and in some cases, longer)
  - Measured 1000BASE-T and 10GBASE-T autonegotiation times are comparable
- Considerations for P802.3bz and the Architecture ad hoc
  - Can 2.5/5GBASE-T autonegotiation and startup times be improved to be more aligned with end-user expectations* and/or requirements?
    - *Assume they will be looking through a 1000BASE-T lens
  - Consider how time-to-link is affected when developing and evaluating 2.5/5GBASE-T autonegotiation proposals

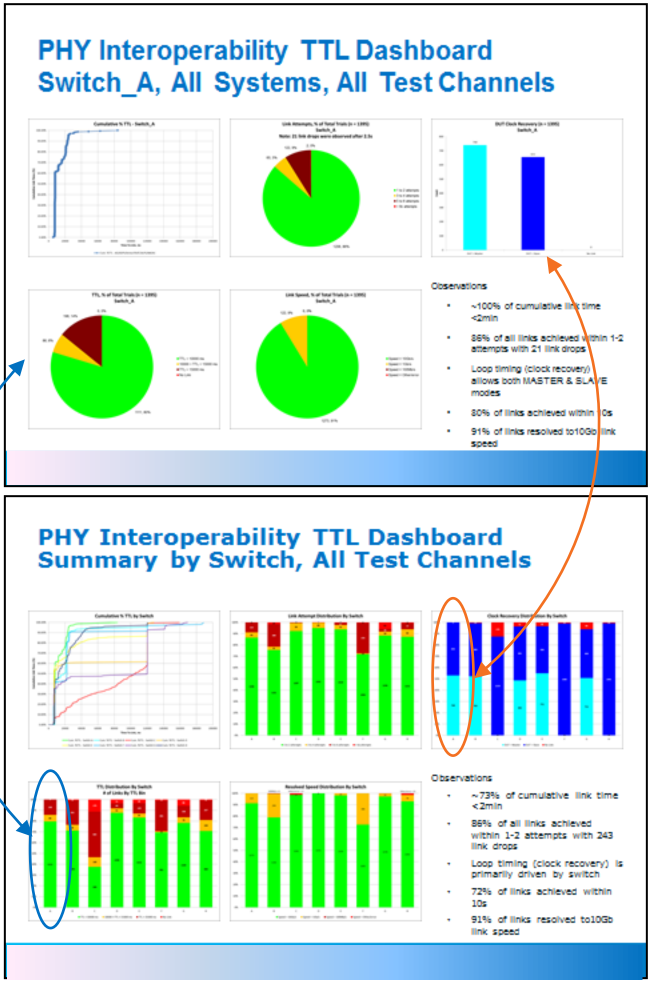
Thank You!

Test Channels

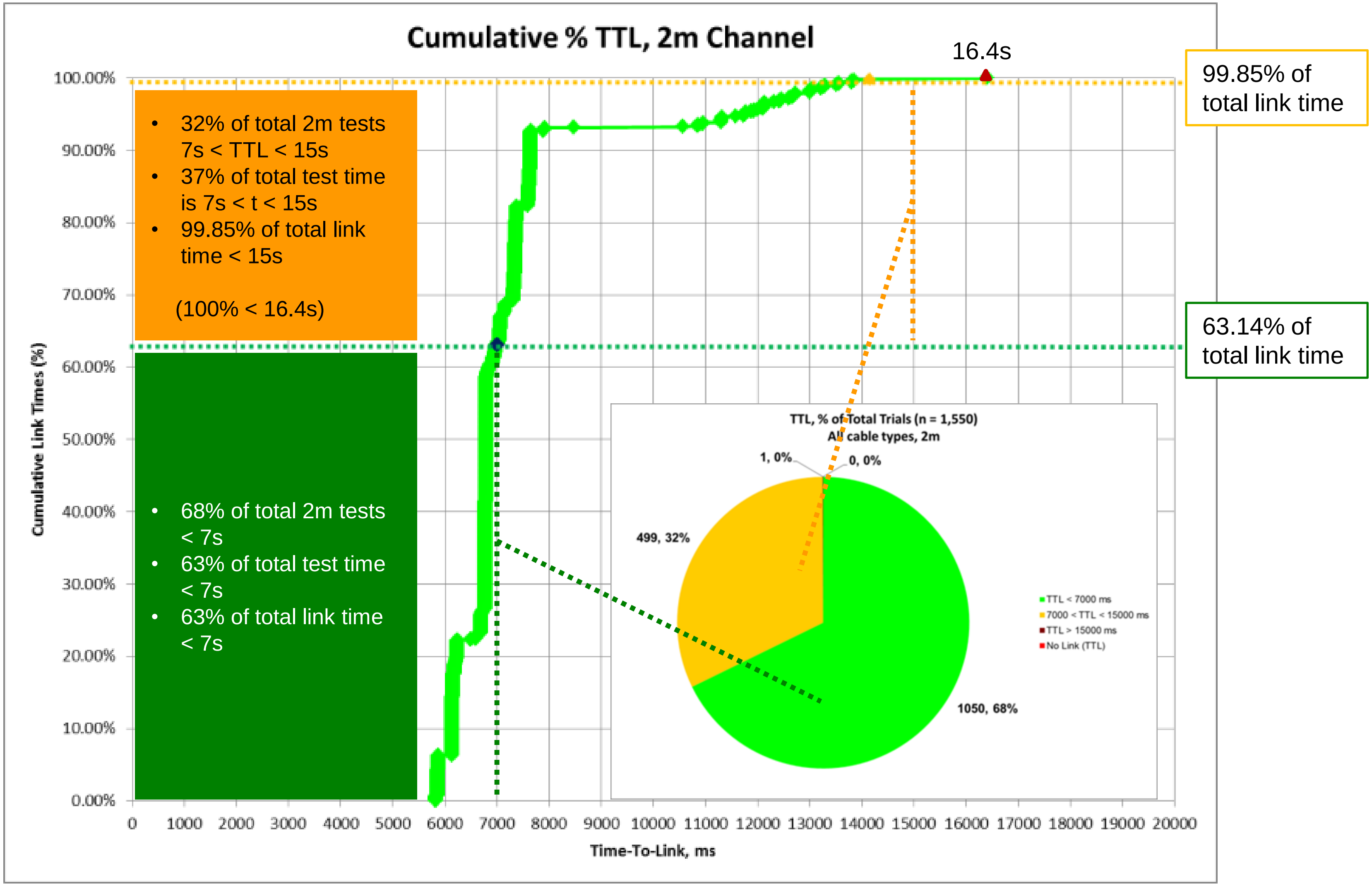
- Focused channel selection using multiple cable types and lengths
  - 2m, 4m, 7m, 30m, 55m, 90m and 100m are “standard” channels for both TTL and BER
  - Other channel lengths (typically 5m increments) are used to check for consistent link behavior over a range of PHY channel solutions (different PBOs, operating margin, delay/delay skew, etc.)
- Includes direct connection, 2-connector, and 4-connector topologies
- Test channel matrix will (of course) be modified for 40GBASE-T

Link Interoperability TTL Dashboards

- Two different dashboard formats are used:
  - "Single dashboards" summarize a single data view
  - "Comparison dashboards" compare multiple "single dashboards" in a slightly different format
- Single dashboard pie charts are represented as a stacked vertical bar charts in comparison dashboards ➔



% TTL (Trials) and Cumulative % TTL (Time) 2m Example

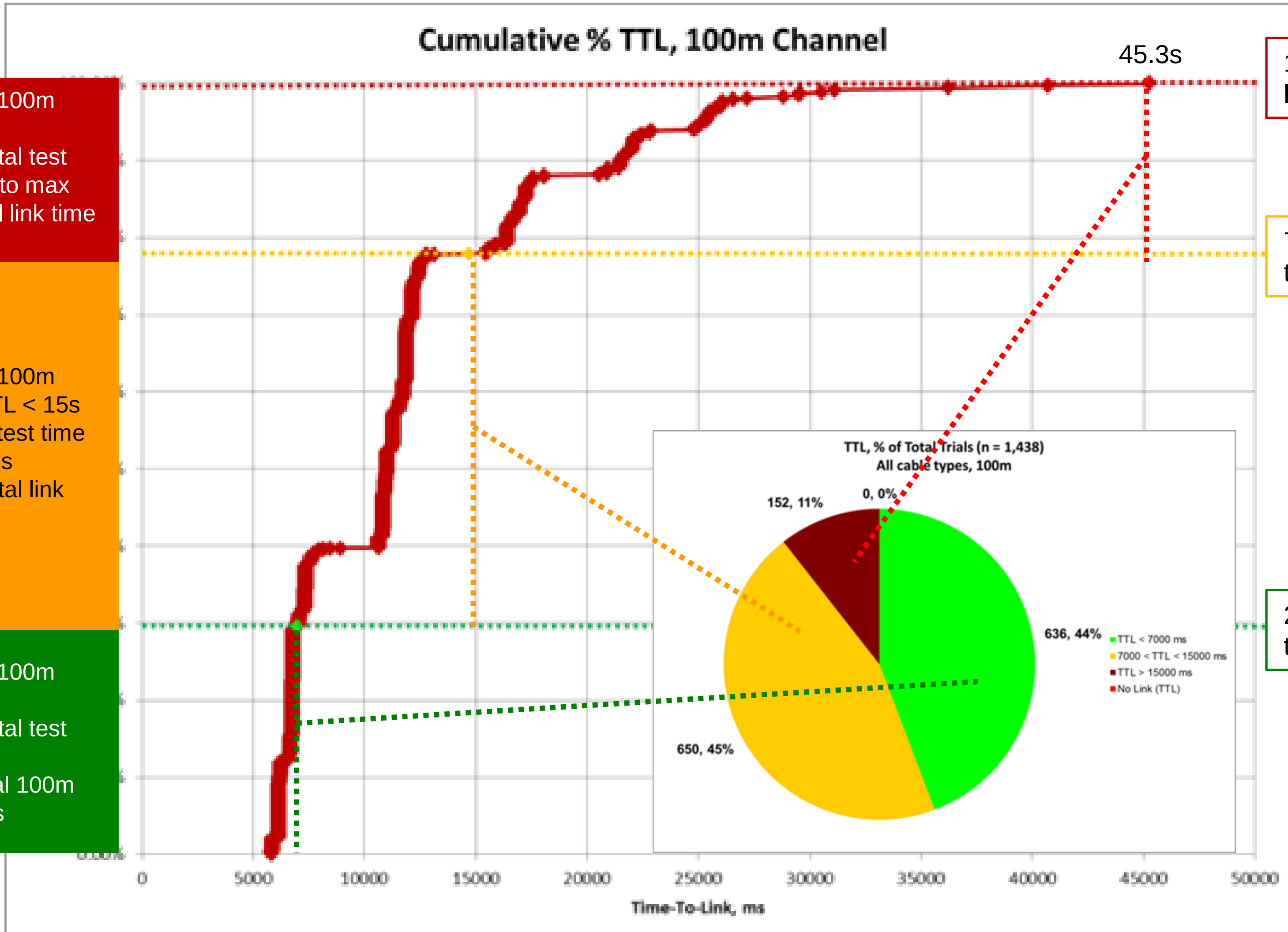


% TTL (Trials) and Cumulative % TTL (Time) 100m Example

- 11% of total 100m tests > 15s
- 29.64% of total test time is >15s to max
- 100% of total link time < 45.3s

- 45% of total 100m tests 7s < TTL < 15s
- 37% of total test time is 7s < t < 15s
- 77.89% of total link time < 15s

- 44% of total 100m tests < 7s
- 29.64% of total test time < 7s
- 29.6% of total 100m link time < 7s



100% of total link time

77.89% of total link time

29.64% of total link time