Karlsruhe Institute of Technology

SPONSORED BY THE

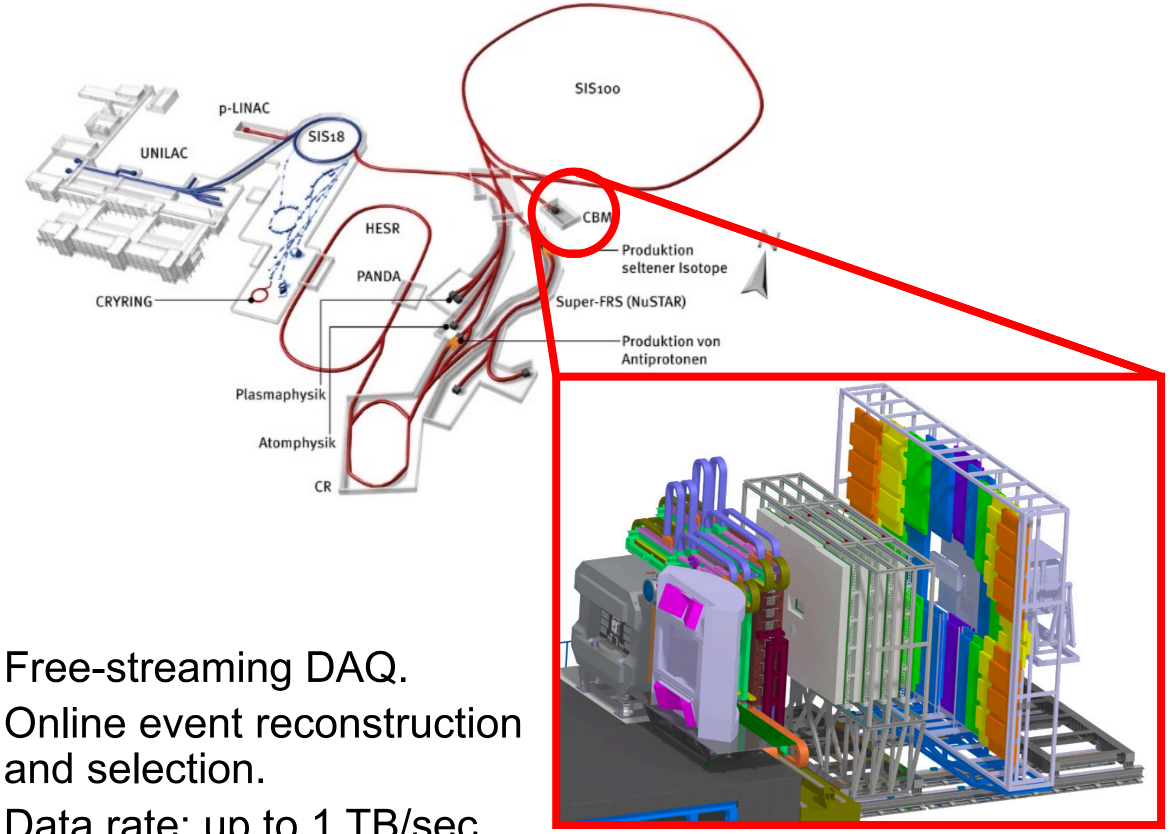
Federal Ministry
of Education
and Research

CBM

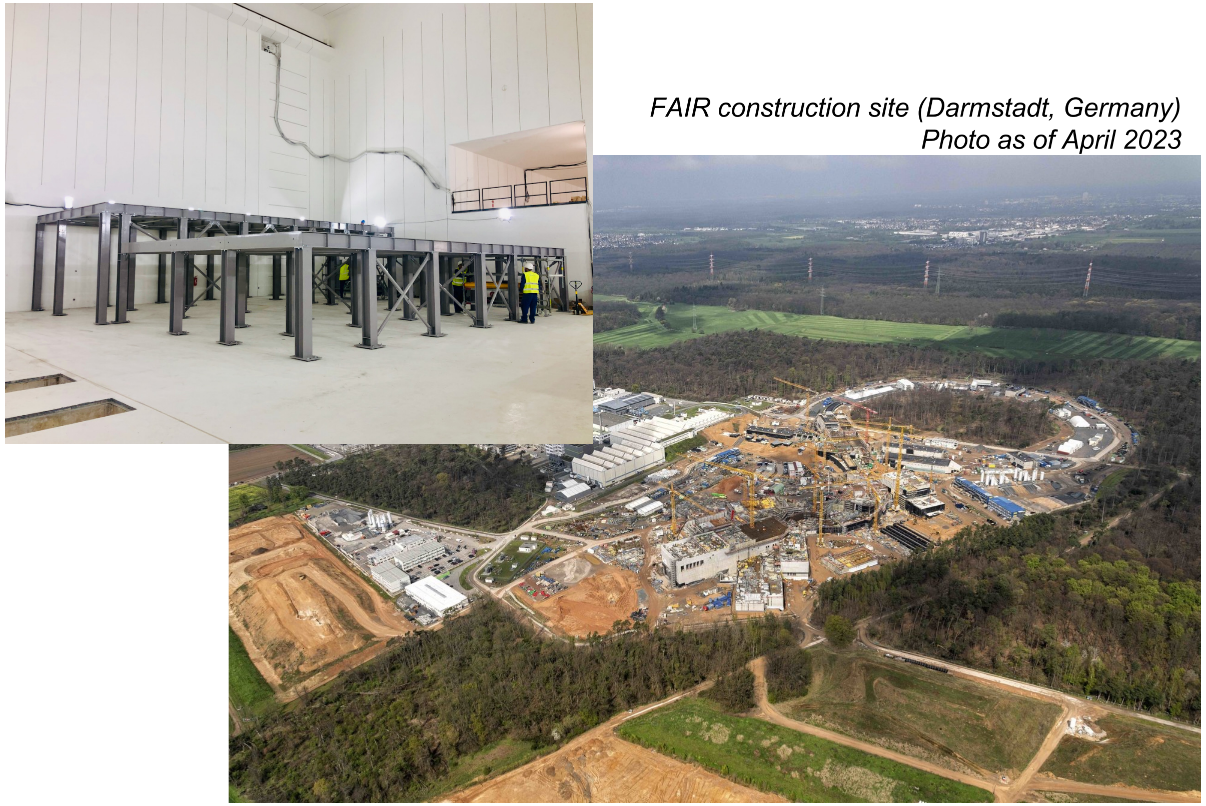
Time and Clock Distribution Over a Hierarchy of Deterministic Optical Links

V. Sidorenko, W. F. J. Müller, W. Zabolotny, I. Fröhlich, D. Emschermann, J. Becker

CBM* experiment overview



- Peak R_{int} is 10 MHz for Au+Au.
- Fast & radiation hard detectors.
- 4D tracking (space, time).



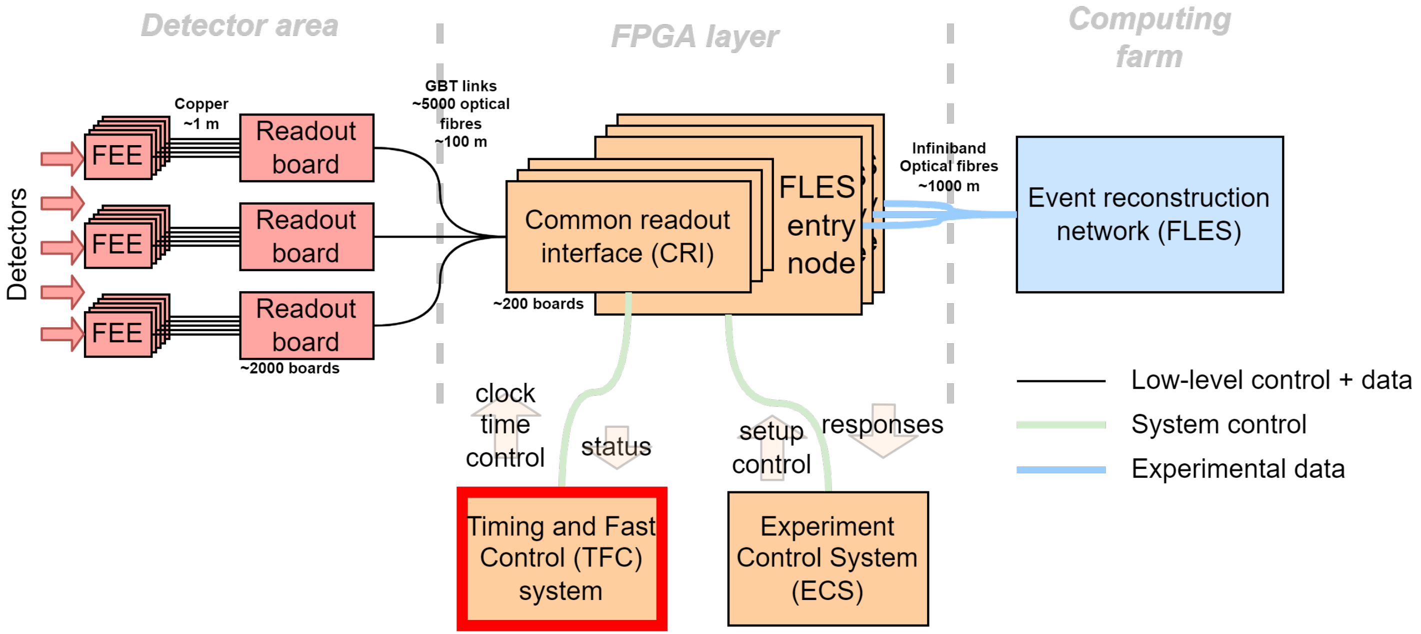
CBM cave. Photo as of May 2023

FAIR construction site (Darmstadt, Germany)
Photo as of April 2023

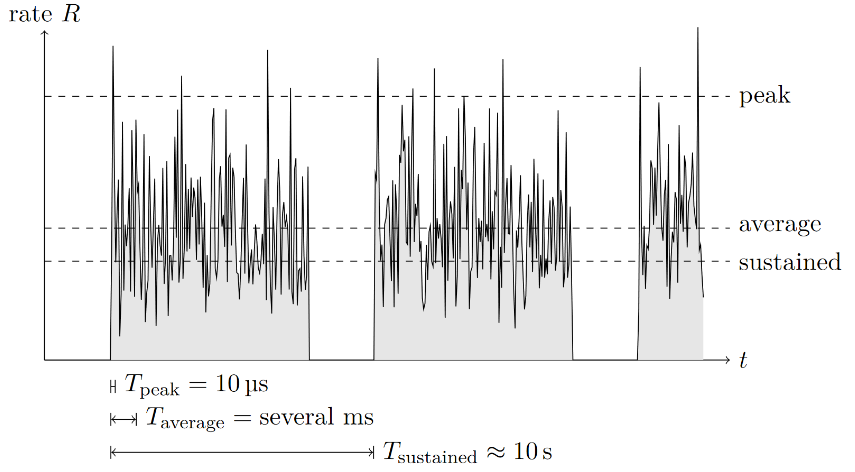
- Free-streaming DAQ.
- Online event reconstruction and selection.
- Data rate: up to 1 TB/sec.

** CBM – Compressed Baryonic Matter*

Streaming data acquisition in CBM



Interaction rate over time



- Distribute a synchronous system-wide clock signal to the CRI endpoints.
- Synchronise the local time counters across the CRI boards.

T F C

Timing Fast control

- Protect the DAQ system from congestion through data throttling.

Requirements for CBM TFC

- Scalability to serve > 200 endpoints with common clock and time.
  - Based on the configuration of the data readout chain.
- < 200 ps synchronization accuracy.
  - Roughly based on timing resolution of RICH, the fastest subsystem apart from ToF.
- < 6 µs fast control response time.
  - Estimation based on the timing constants in the readout system and throttling strategy simulations*.

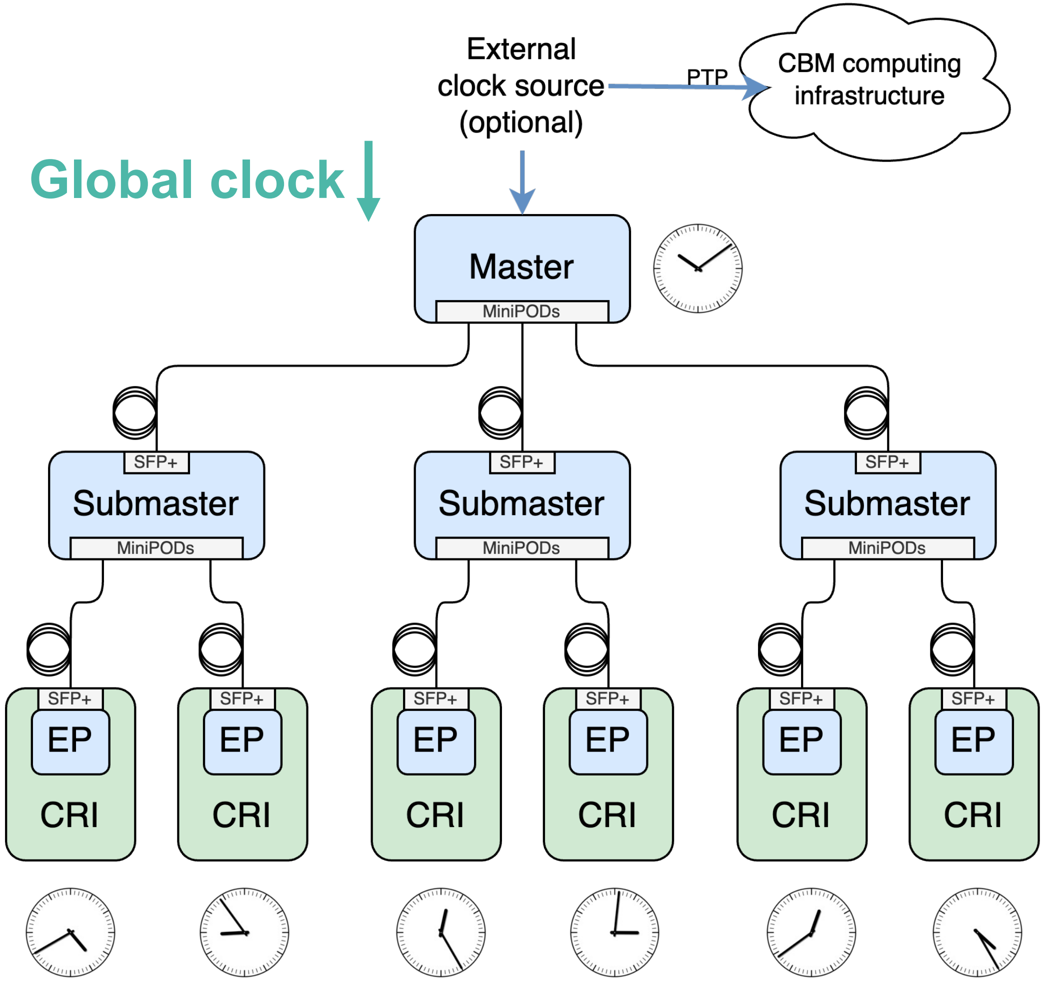
** X. Gao, D. Emschermann, J. Lehnert, and W. F. J. Müller, "Throttling strategies and optimization of the trigger-less streaming DAQ system in the CBM experiment," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 978. Elsevier BV, p. 164442, Oct. 2020. doi: 10.1016/j.nima.2020.164442.*

TFC concept

T F C

Timing

- The Master node receives the external clock reference.
- Time counter in the Master node defines 64-bit experiment-wide TFC time.
- TFC time is initialized via PCIe control interface.



CRI local time

EP – Endpoint
CRI – Common Readout Interface

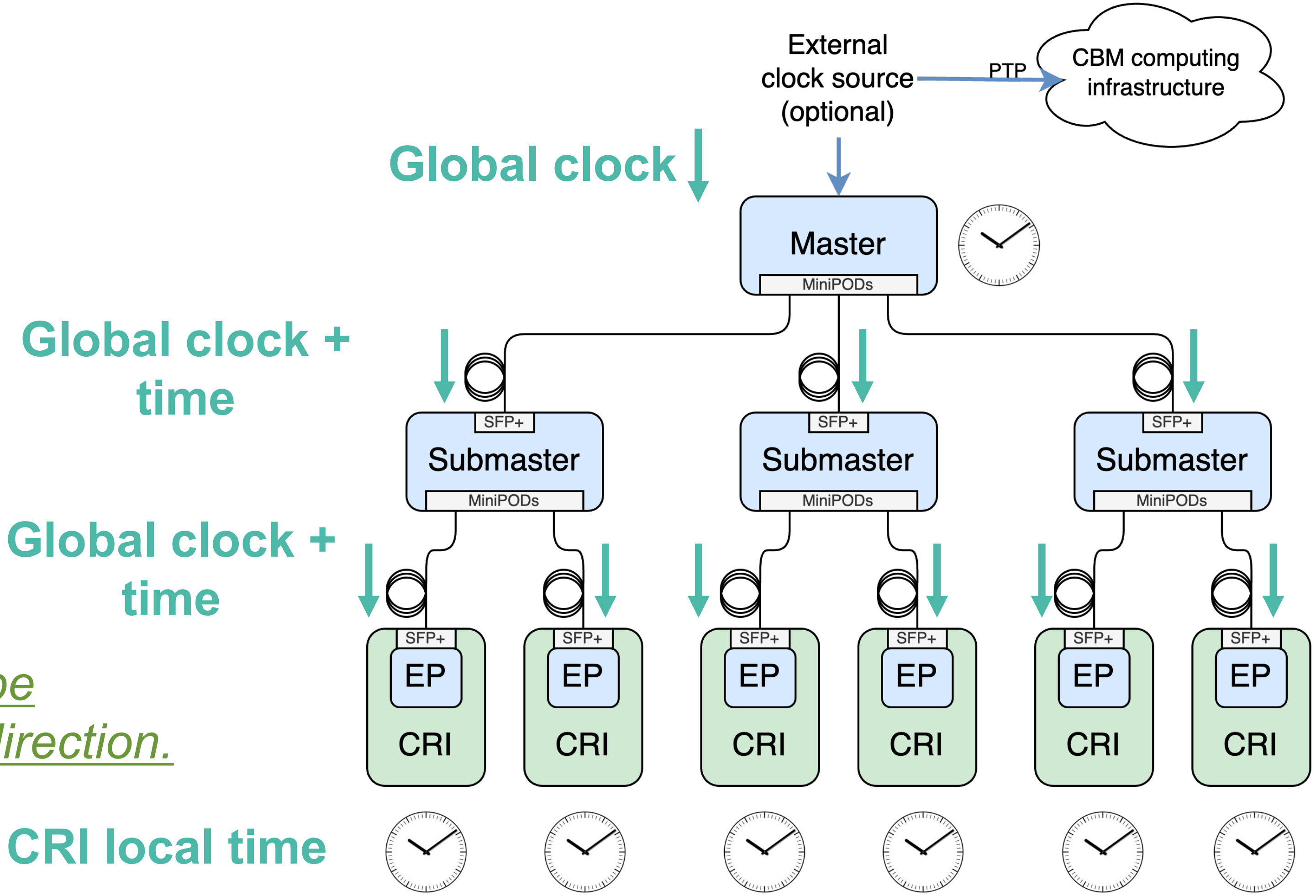
TFC concept

T F C

Timing

- TFC time is distributed over a hierarchy of optical links.
- Global clock is distributed over the same links.
- Intermediate nodes recover clock from upstream links and reuse it for further communication.

Latency of the optical links must be deterministic in the downstream direction.



EP – Endpoint
CRI – Common Readout Interface

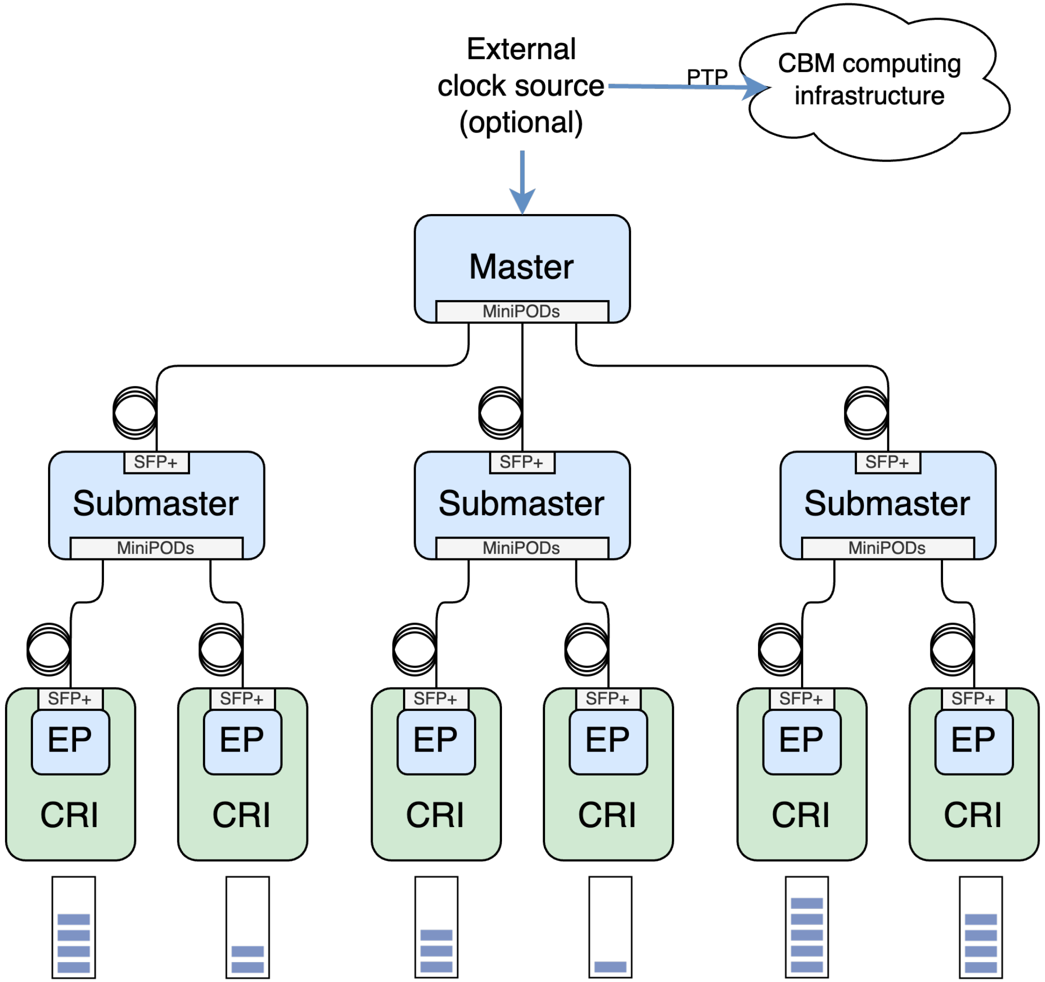
TFC concept

T F C

Fast control

- Each CRI board issues status information on FIFO fill level.
- The status information is aggregated and passed to the TFC Master.

TFC links must be bidirectional.



FEE buffer fill level

EP – Endpoint
CRI – Common Readout Interface

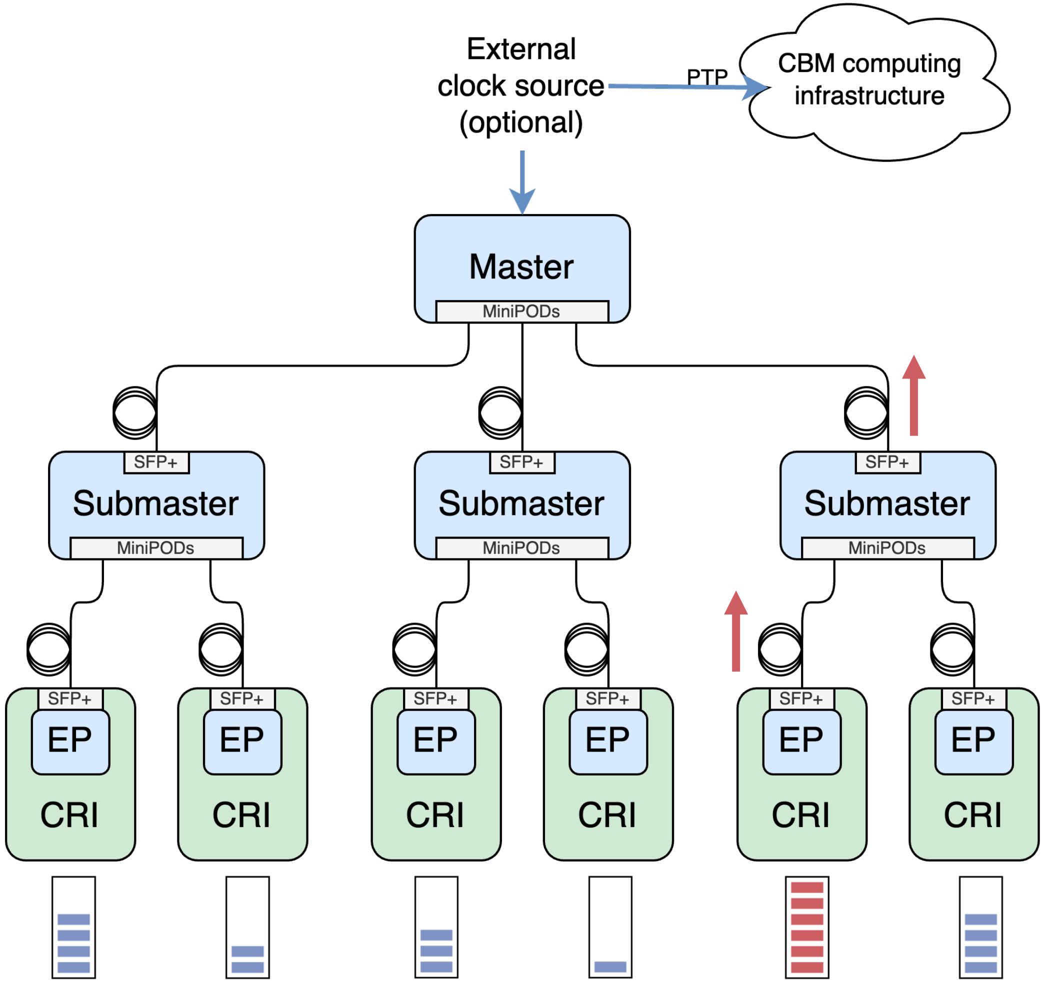
TFC concept

T **F C**

Fast control

- When the buffers are dangerously occupied, this information is propagated to the Master node.
- With the upstream link ratio of 47:1, Submasters must aggregate the data.

TFC links must have low latency.



FEE buffer fill level

EP – Endpoint
CRI – Common Readout Interface

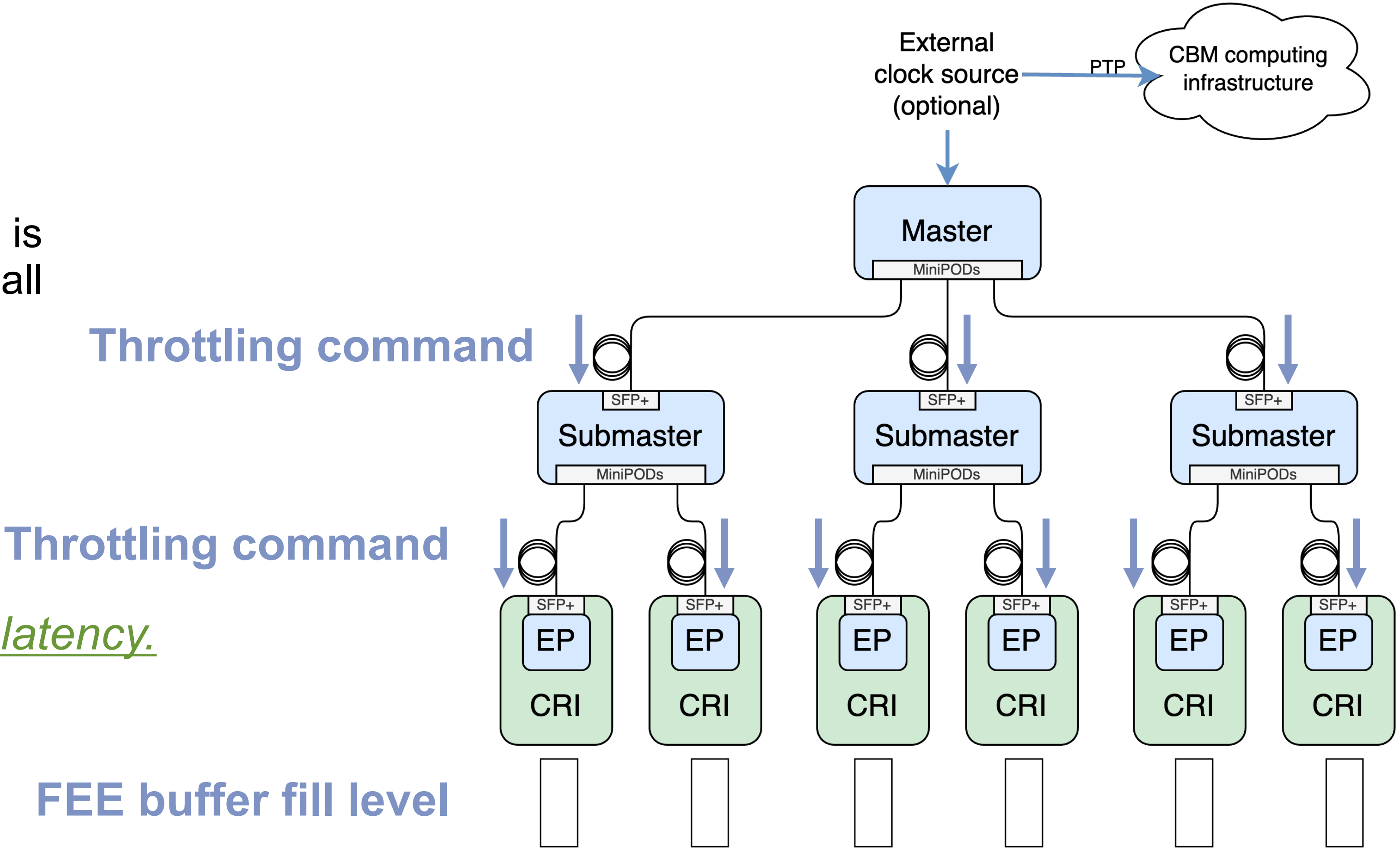
TFC concept

T F C

Fast control

- Data throttling command is issued and broadcast to all Endpoints.
- Data load on buffers is relieved to prevent uncontrolled event loss.

TFC links must have low latency.



EP – Endpoint
CRI – Common Readout Interface

TFC architecture

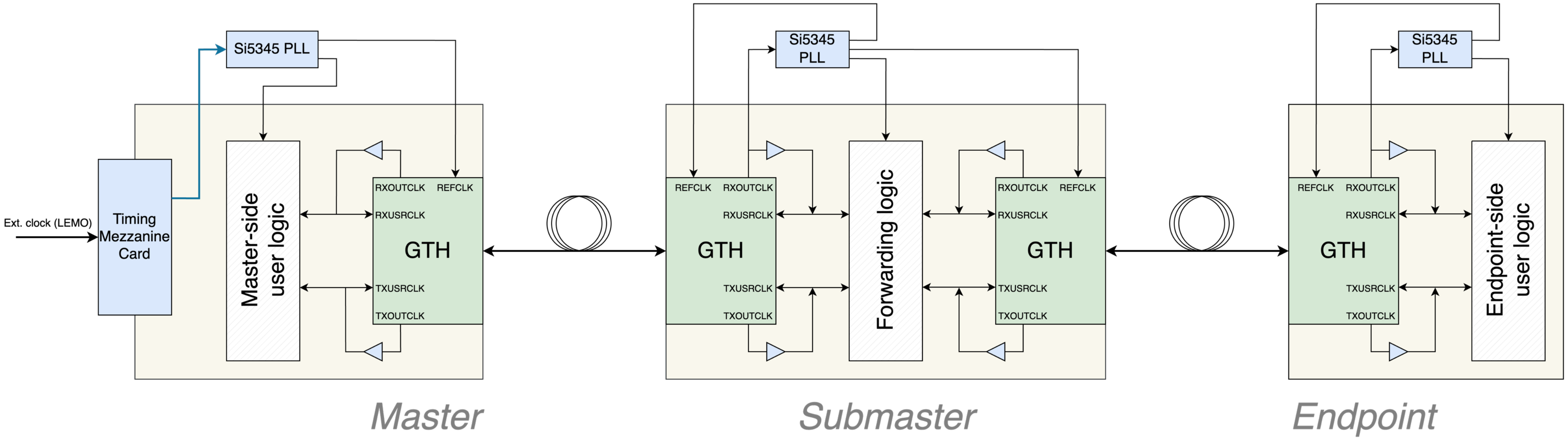
- Platform board: BNL-712

Mezzanine cards:

- Master - WR TMC
- Endpoint - TTC-PON TMC

- Board highlights:
  - Developed by Brookhaven National Lab, USA
  - Xilinx XCKU115FLVF1924-2E FPGA
  - Si5345 jitter cleaner for recovered clock
  - 48 optical connections (1 SFP, 47 Broadcom MiniPOD)
  - PCIe Gen3 x16 lane interface

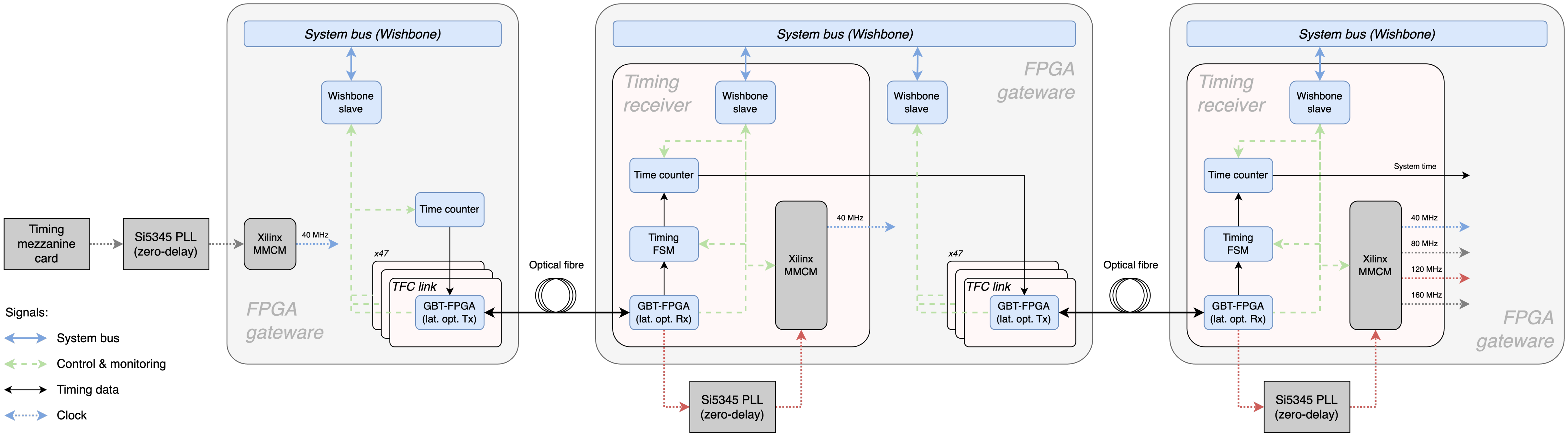
TFC architecture



- Clock signal is embedded into the data communication.
- Each node recovers clock from an upstream link and uses it for its own logic and for further downstream links if needed.
- All components in the clock cascading chain have a deterministic input-to-output delay.

TFC architecture

- Wishbone as the system bus with AGWB* infrastructure.
- GBT-FPGA provides latency-deterministic communication over fibre.



* *W. M. Zabołotny, M. Gumiński, M. Kruszewski, and W. F. J. Müller, "Control and Diagnostics System Generator for Complex FPGA-Based Measurement Systems," Sensors, vol. 21, no. 21. MDPI AG, p. 7378, Nov. 06, 2021. doi: 10.3390/s21217378.*

Clock jitter and skew in the system

Latency determinism over one hop has been previously evaluated*.

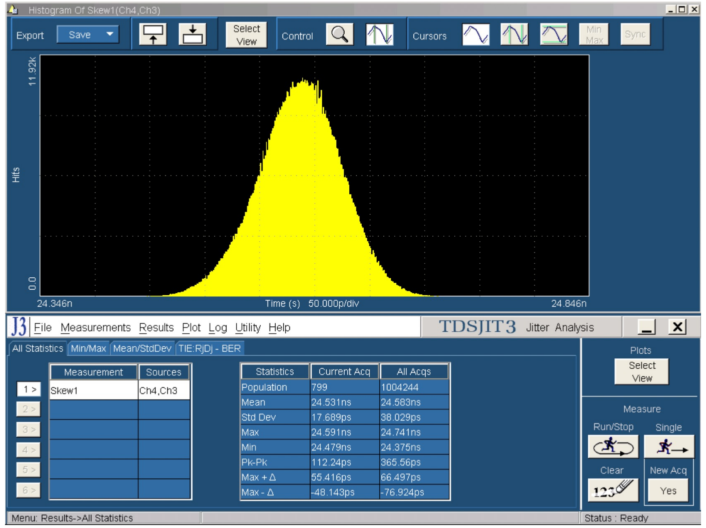
It is still unclear how the timing error will scale in a larger system.

Goals of the current study:

- Evaluate clock jitter in the system nodes.
- Estimate how clock jitter changes with added network layers and endpoint nodes.
- Estimate how clock skew changes with added network layers and endpoint nodes.

Hardware used:

- 3x BNL-712 + SFP mezzanines.
- Tektronix TDS6154C (4 ch, 15 GHz, 40 GSa/s) + TDSJIT3 Advanced jitter measurement app.



	Measurement	Sources
1 >	Skew1	Ch4,Ch3
2 >		
3 >		
4 >		
5 >		
6 >		

Statistics	Current Acq	All Acqs
Population	799	1004244
Mean	24.531ns	24.583ns
Std Dev	17.689ps	38.029ps
Max	24.591ns	24.741ns
Min	24.479ns	24.375ns
Pk-Pk	112.24ps	365.56ps
Max + Δ	55.416ps	66.497ps
Max - Δ	-48.143ps	-76.924ps

** V. Sidorenko, W. F. J. Müller, W. Zabolotny, I. Fröhlich, D. Emschermann, and J. Becker, “Evaluation of GBT-FPGA for timing and fast control in CBM experiment,” Journal of Instrumentation, vol. 18, no. 02. IOP Publishing, p. C02052, Feb. 01, 2023. doi: 10.1088/1748-0221/18/02/c02052.*

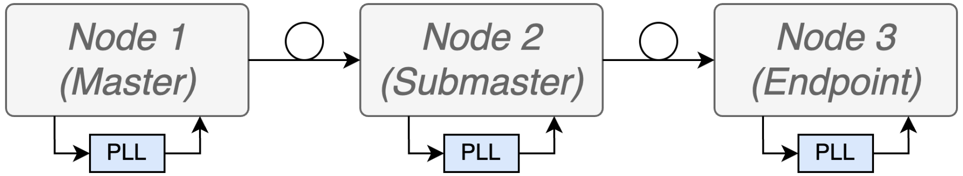
Test conditions

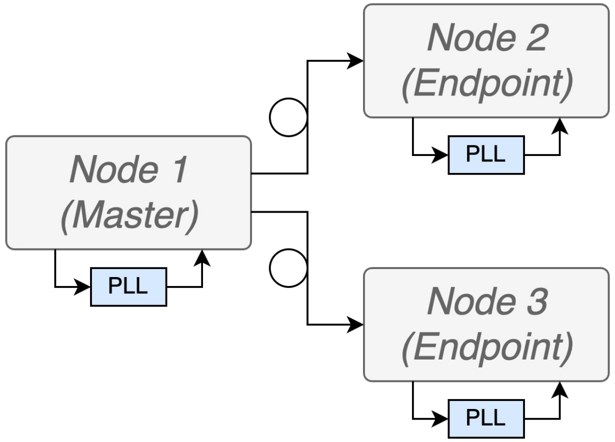
- Direct measurement of the 40 MHz system clock.
- Clock recovery with Silabs Si5345 (Rev B) at ~87 Hz loop bandwidth.
- Air-conditioned room with insignificant temperature variation.

Measurements:

- Jitter measurement on each node.
- Clock skew between nodes at each hop.

- 3 measurements, >1M samples each
- Two setup configurations:





Test results: 2 hop configuration

Master period jitter

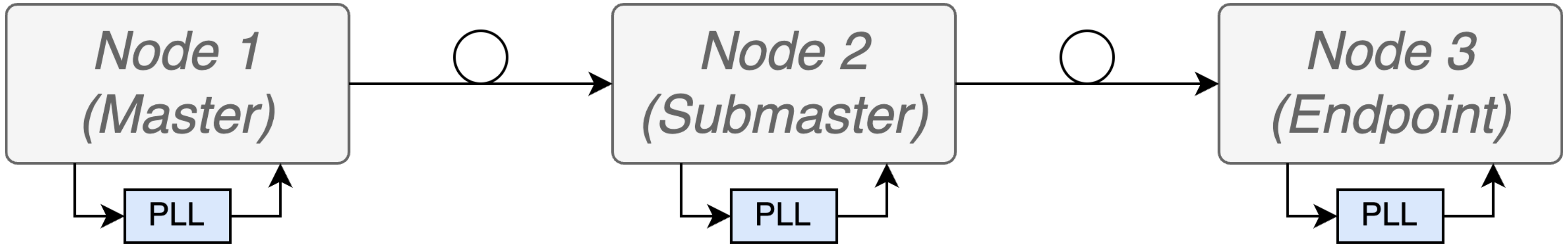
Test no.	μ, ns	σ, ps
1	25.001	9.696
2	25.001	9.893
3	25.001	9.902
AVG	25.001	9.831

Submaster period jitter

Test no.	μ, ns	σ, ps
1	25.001	9.691
2	25.001	9.524
3	25.001	9.616
AVG	25.001	9.610

Endpoint period jitter

Test no.	μ, ns	σ, ps
1	25.001	12.684
2	25.001	10.182
3	25.001	10.527
AVG	25.001	11.131



Master-Submaster skew

Test no.	μ, ns	σ, ps
1	13.756	24.377
2	14.298	21.311
3	14.284	21.527
AVG	14.113	25.738

Submaster-Endpoint skew

Test no.	μ, ns	σ, ps
1	11.655	27.840
2	11.638	25.955
3	11.688	30.736
AVG	11.660	28.177

Master-Endpoint skew

Test no.	μ, ns	σ, ps
1	0.417	38.029
2	0.931	33.191
3	0.978	37.407
AVG	0.775	36.209

Test results: 2 hop configuration

Master period jitter

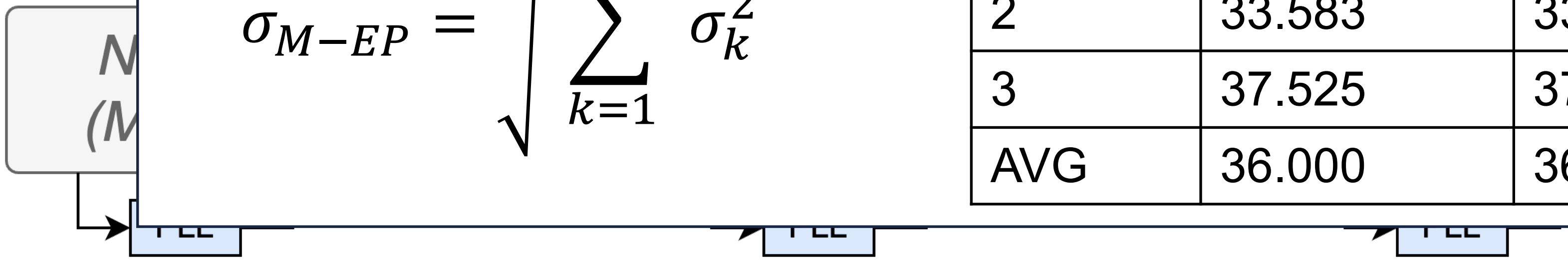
Test no.	μ, ns	σ, ps
1		
2		
3		
AVG		

Submaster period jitter

Test no.	μ, ns	σ, ps

Endpoint period jitter

Test no.	μ, ns	σ, ps



Hypothesis 1:

$$\sigma_{M-EP} = \sqrt{\sum_{k=1}^{N_{hops}} \sigma_k^2}$$

Test no.	σ calculated	σ measured
1	37.004	38.029
2	33.583	33.191
3	37.525	37.407
AVG	36.000	36.209

Master-Submaster skew

Test no.	μ, ns	σ, ps
1	13.756	24.377
2	14.298	21.311
3	14.284	21.527
AVG	14.113	22.405

Submaster-Endpoint skew

Test no.	μ, ns	σ, ps
1	11.655	27.840
2	11.638	25.955
3	11.688	30.736
AVG	11.660	28.177

Master-Endpoint skew

Test no.	μ, ns	σ, ps
1	0.417	38.029
2	0.931	33.191
3	0.978	37.407
AVG	0.775	36.209

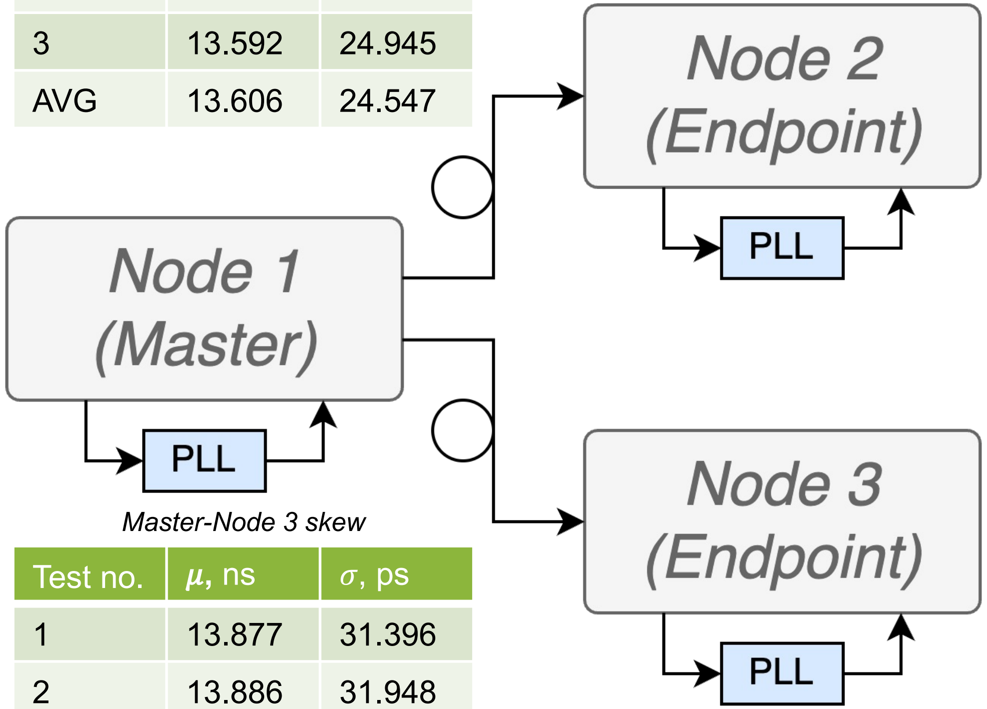
Test results: 2 endpoint configuration

Master-Node 2 skew

Test no.	μ, ns	σ, ps
1	13.582	24.593
2	13.645	24.102
3	13.592	24.945
AVG	13.606	24.547

Master period jitter

Test no.	μ, ns	σ, ps
1	25.001	9.911
2	25.001	9.917
3	25.001	10.057
AVG	25.001	9.962



Master-Node 3 skew

Test no.	μ, ns	σ, ps
1	13.877	31.396
2	13.886	31.948
3	13.956	34.212
AVG	13.906	32.519

Node 2 period jitter

Test no.	μ, ns	σ, ps
1	25.001	12.997
2	25.001	14.497
3	25.001	14.411
AVG	25.001	13.968

Node 3 period jitter

Test no.	μ, ns	σ, ps
1	25.001	8.538
2	25.001	10.400
3	25.001	7.836
AVG	25.001	8.925

Node 2-Node 3 skew

Test no.	μ, ns	σ, ps
1	0.293	30.869
2	0.241	31.482
3	0.368	34.050
AVG	0.301	32.134

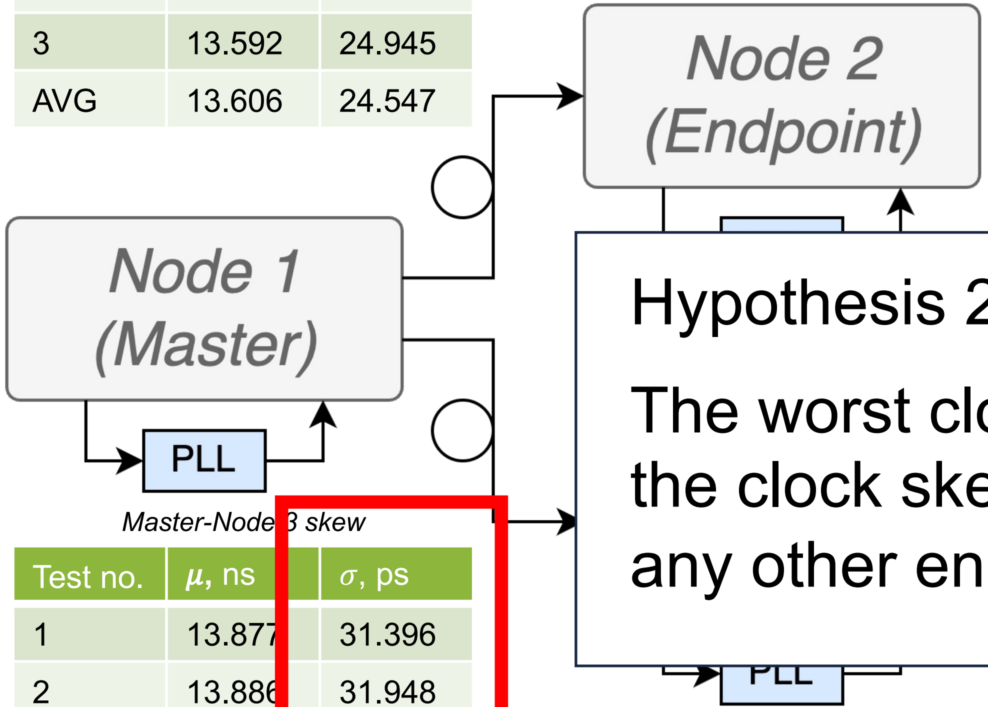
Test results: 2 endpoint configuration

Master-Node 2 skew

Test no.	μ, ns	σ, ps
1	13.582	24.593
2	13.645	24.102
3	13.592	24.945
AVG	13.606	24.547

Master period jitter

Test no.	μ, ns	σ, ps
1	25.001	9.911
2	25.001	9.917
3	25.001	10.057
AVG	25.001	9.962



Node 2 period jitter

Test no.	μ, ns	σ, ps
1	25.001	12.997
2	25.001	14.497
3	25.001	14.411
AVG	25.001	13.968

Master-Node 3 skew

Test no.	μ, ns	σ, ps
1	13.877	31.396
2	13.886	31.948
3	13.956	34.212
AVG	13.906	32.519

Node 2-Node 3 skew

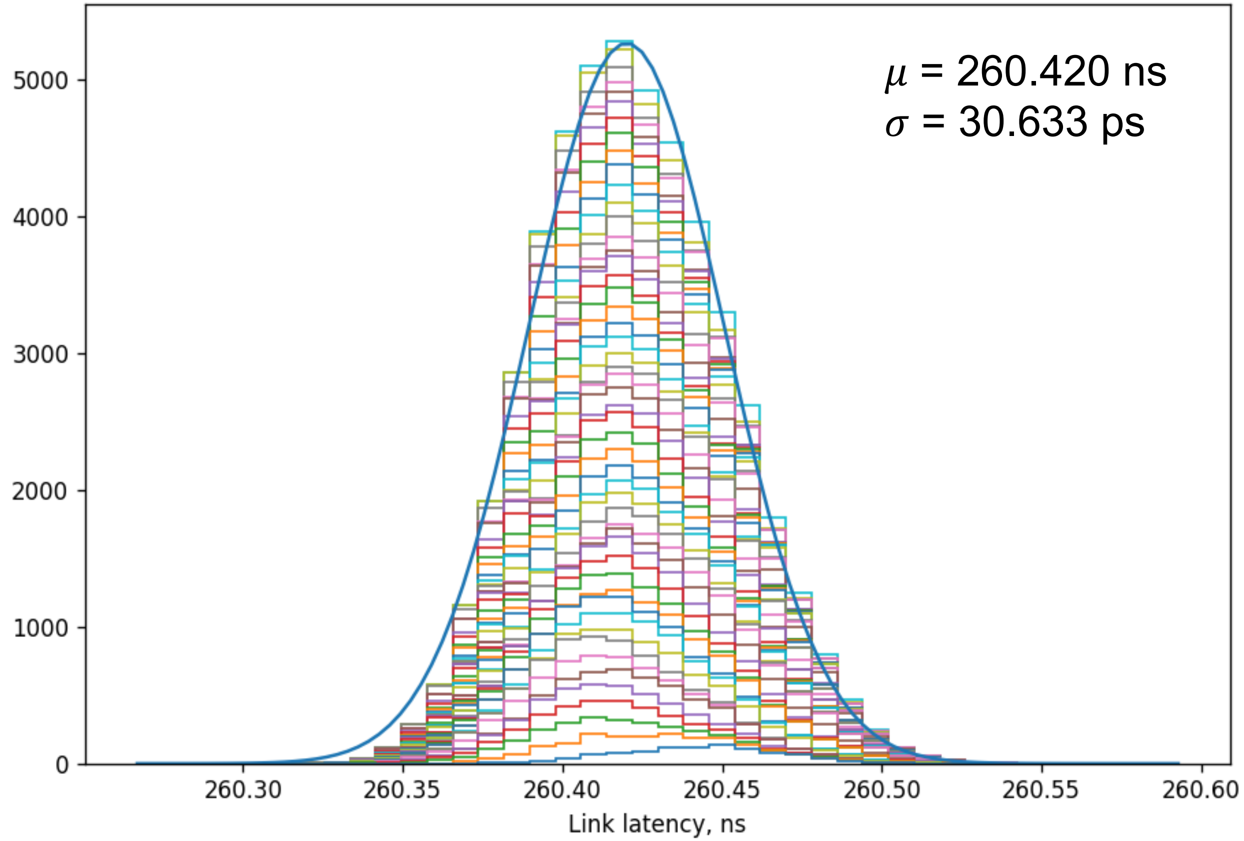
Test no.	μ, ns	σ, ps
	0.293	30.869
	0.241	31.482
	0.368	34.050
	0.301	32.134

Hypothesis 2:

The worst clock skew σ defines the clock skew σ between this and any other endpoint in the system.

Applying the results to the TFC system

Link latency measurements.
50 runs, 1000 samples each (full power cycle between runs).



- With the current hardware platform, 2 hops are required to serve 200 CRI boards.
- Clock skew σ over 2 hops:

$$\sigma_{M-EP} = \sigma\sqrt{2} \approx 43.322\, ps$$

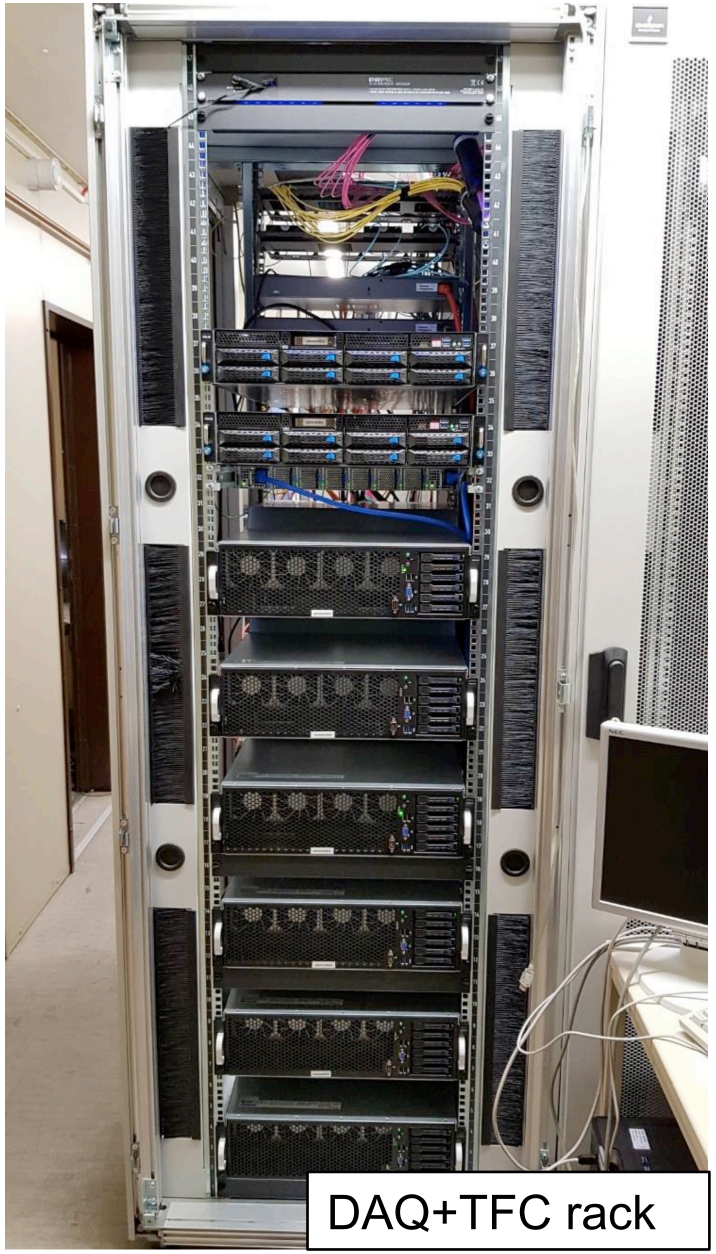
- Before hypothesis 2 can be applied, the worst-case link must be identified.

Conclusions

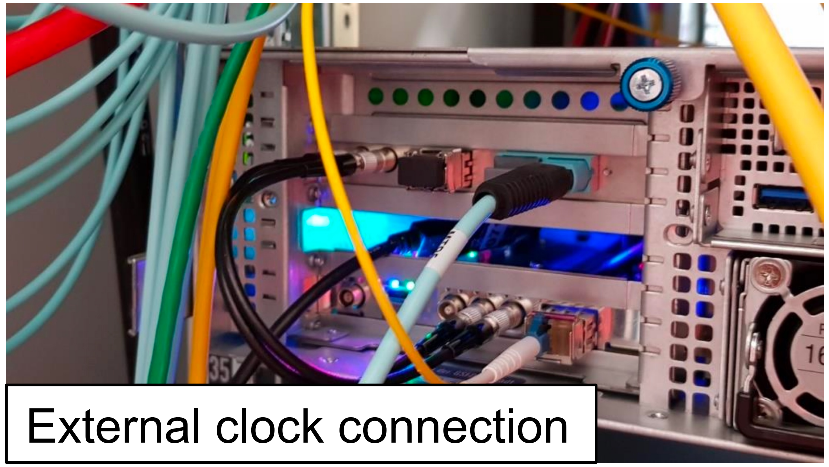
- Vertical scaling (adding layers or hops) appears to be predictable.
- Horisontal scaling appears to be defined by the worst-case link and requires its identification.

- Insight has been gained into how timing distribution error scales with adding network nodes and layers
- …although there is more insight to gain!

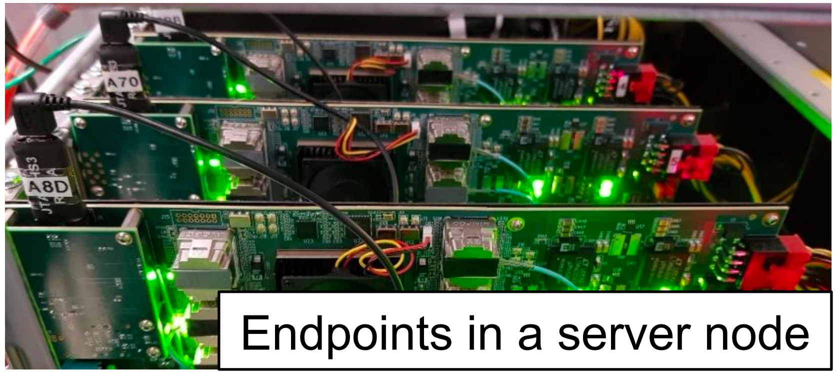
- Although more accurate estimations have yet to be done, performance of the timing distribution system looks very promising for the needs of the experiment.



DAQ+TFC rack



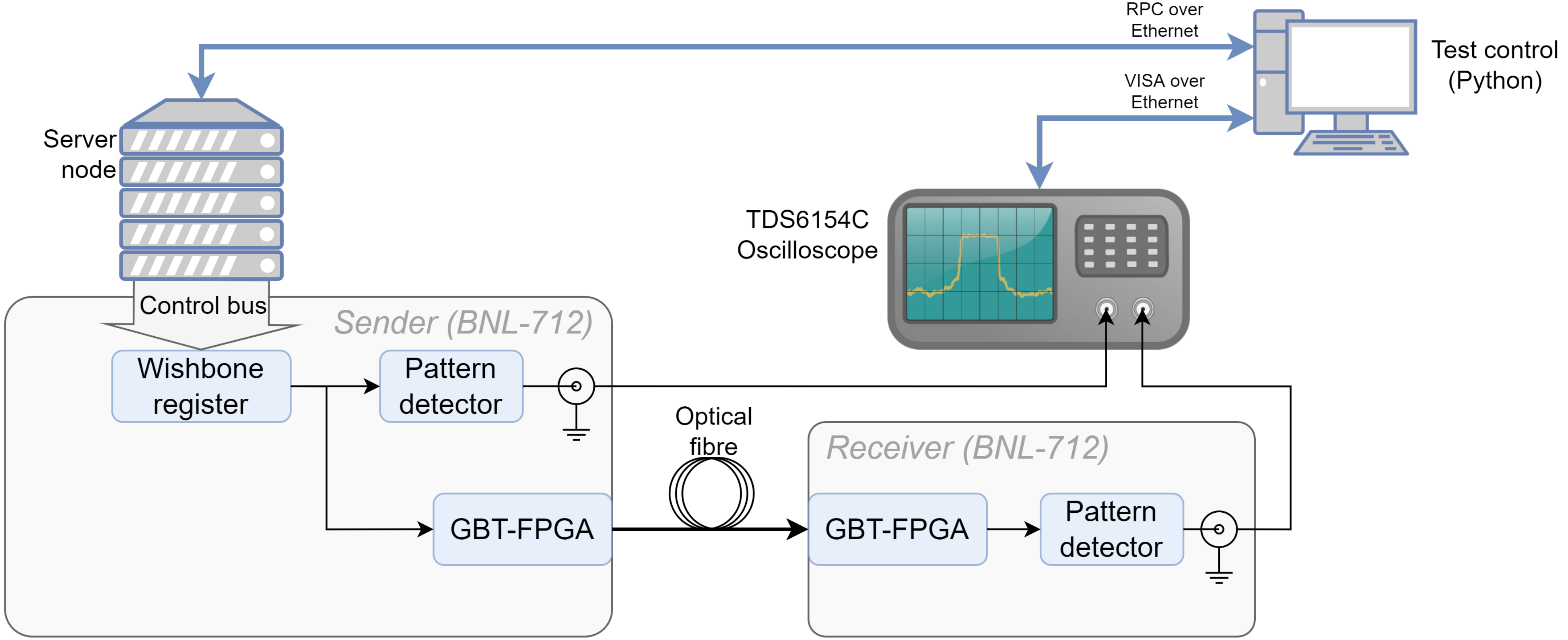
External clock connection

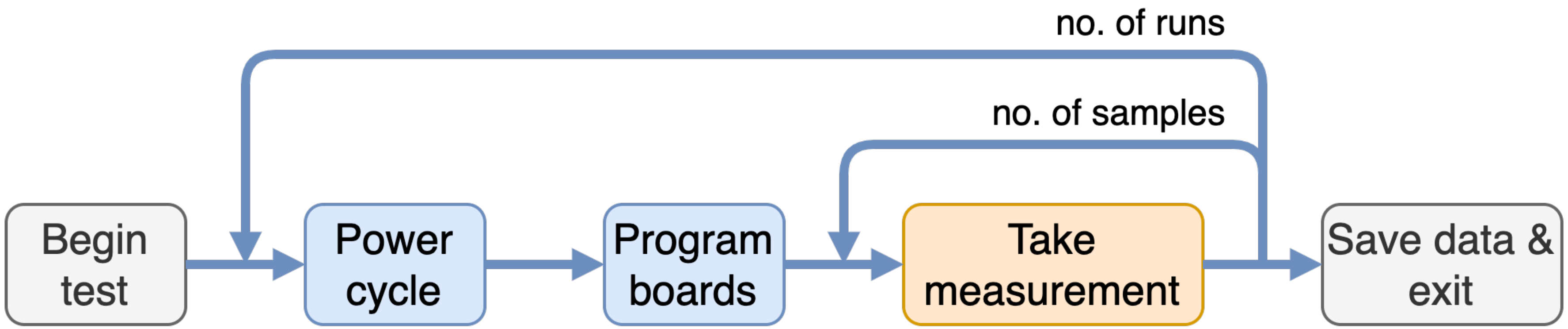


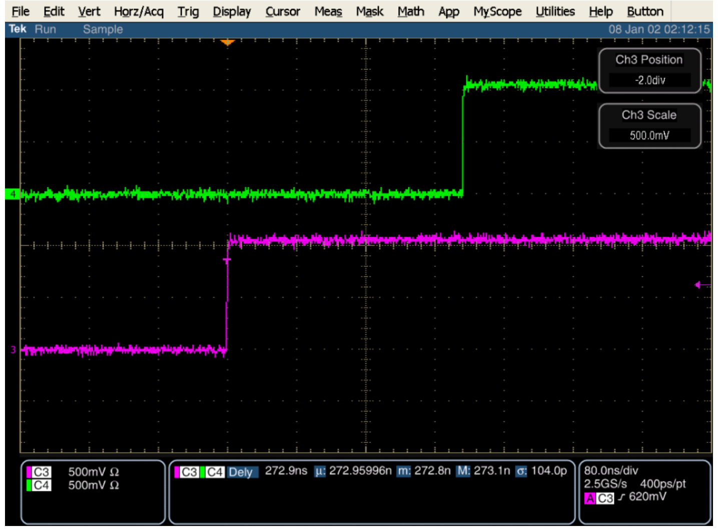
Endpoints in a server node

Thank you!

Backup slides


RPC over Ethernet
VISA over Ethernet
Test control (Python)
Server node
TDS6154C Oscilloscope
Control bus
Sender (BNL-712)
Wishbone register
Pattern detector
Optical fibre
Receiver (BNL-712)
GBT-FPGA
GBT-FPGA
Pattern detector





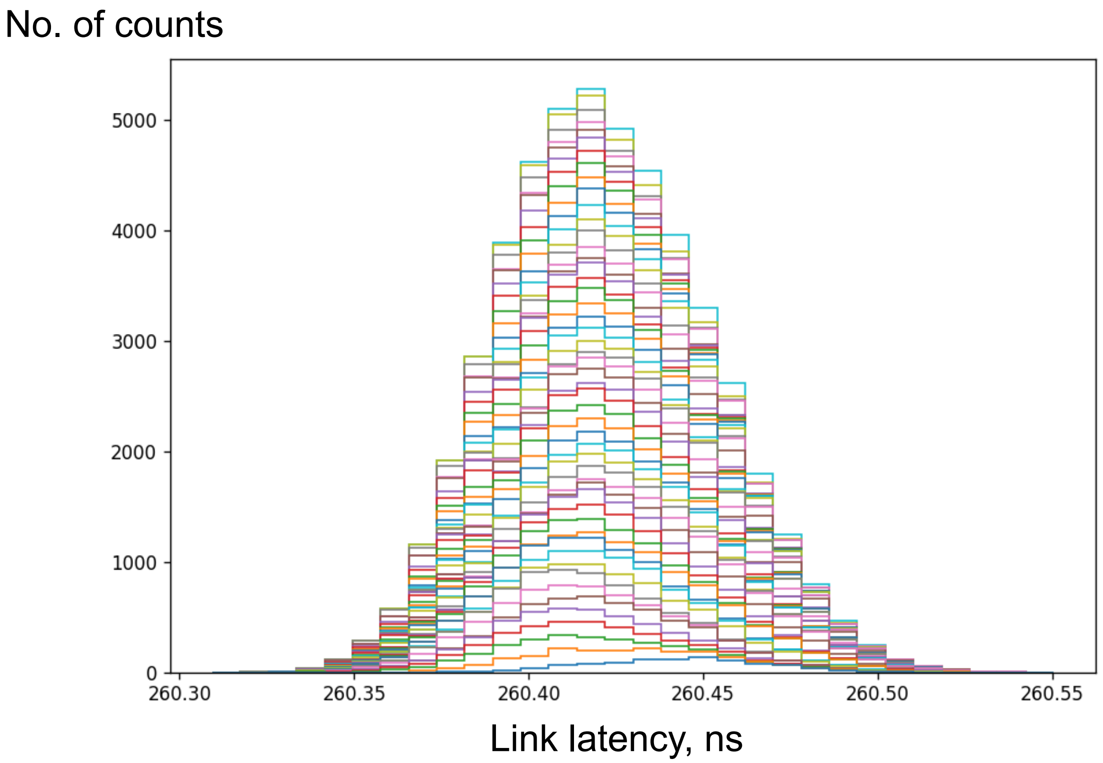
```
Run      Sample                                08 Jan 02 02:12:15
Ch3 Position  -2.0div
Ch3 Scale     500.0mV
C3  500mV Ω
C4  500mV Ω
C3 C4 Dely  272.9ns  µ: 272.95996n  m: 272.8n  M: 273.1n  σ: 104.0p
80.0ns/div
2.5GS/s    400ps/pt
A C3 ↗ 620mV
```

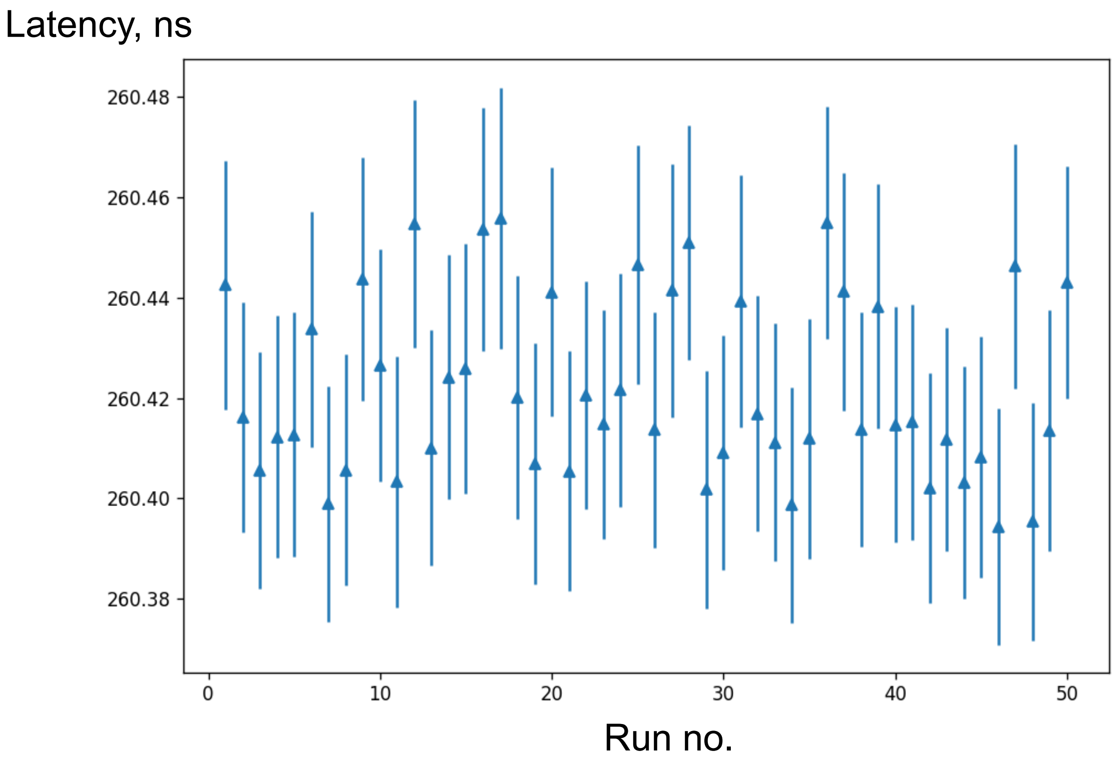
An example of a measurement (sample)

- Measurement speed could be better:
  - collecting 1k samples takes ~70 min
  - ~11 hours for 10 runs

- *Sample* – one instance of latency measurement on the link.
- *Run* – a set of consecutive samples.

50 runs, 1000 samples each (full power cycle between runs)


No. of counts
Link latency, ns


Latency, ns
Run no.