SYNC in 5G and AI

Satheesh Kumar S, HPE
Principal Engineer

Juniper Networks | Driven by Experience

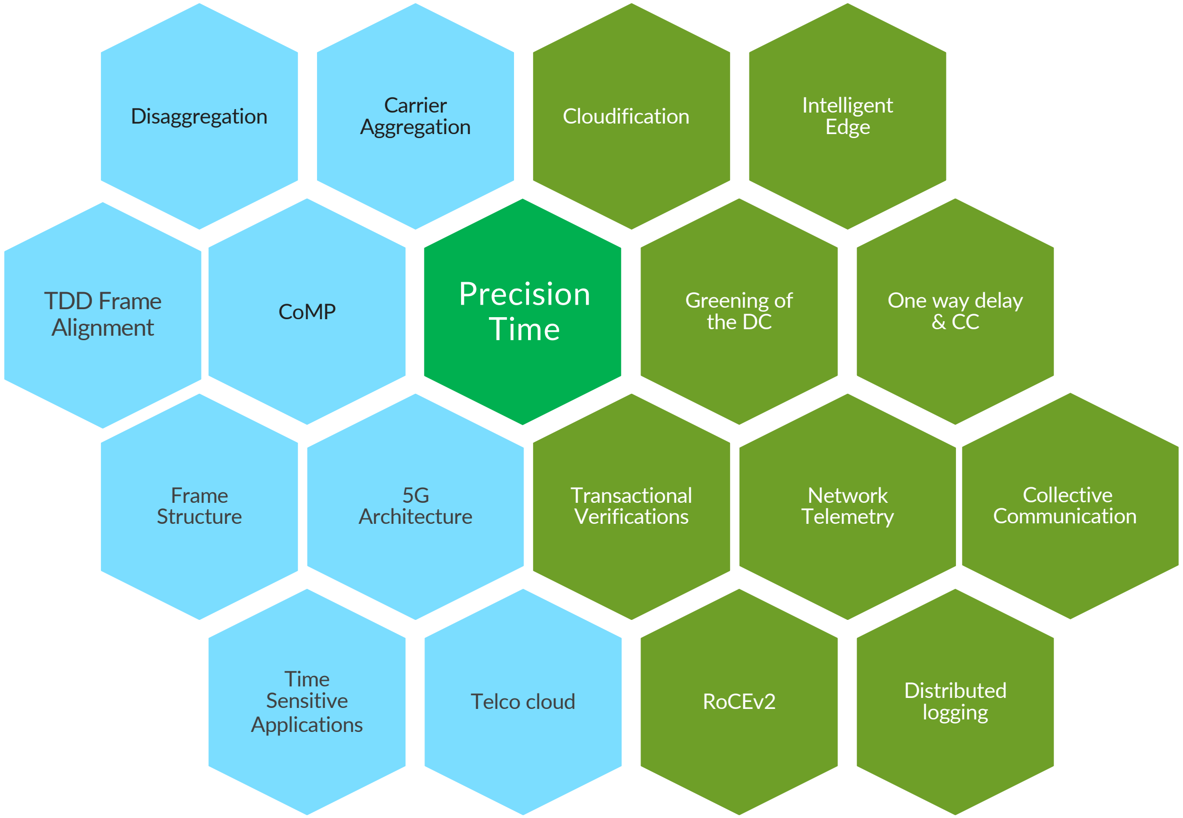
Agenda

- Sync in 5G and AI
- 5G Timing Requirements
- 5G Timing Overview
- O-RAN Configurations and Limit
- AI for Networking and Networking for AI
- Timing in AI Cluster
- TAP DC Profile
- AI Cluster Timing Requirements
- Predictive Modelling
- Summary

As resource becomes more efficient to use, the cost of using the resource drops, which can stimulate demand.

-William Stalling Jevons

SYNC in 5G and AI

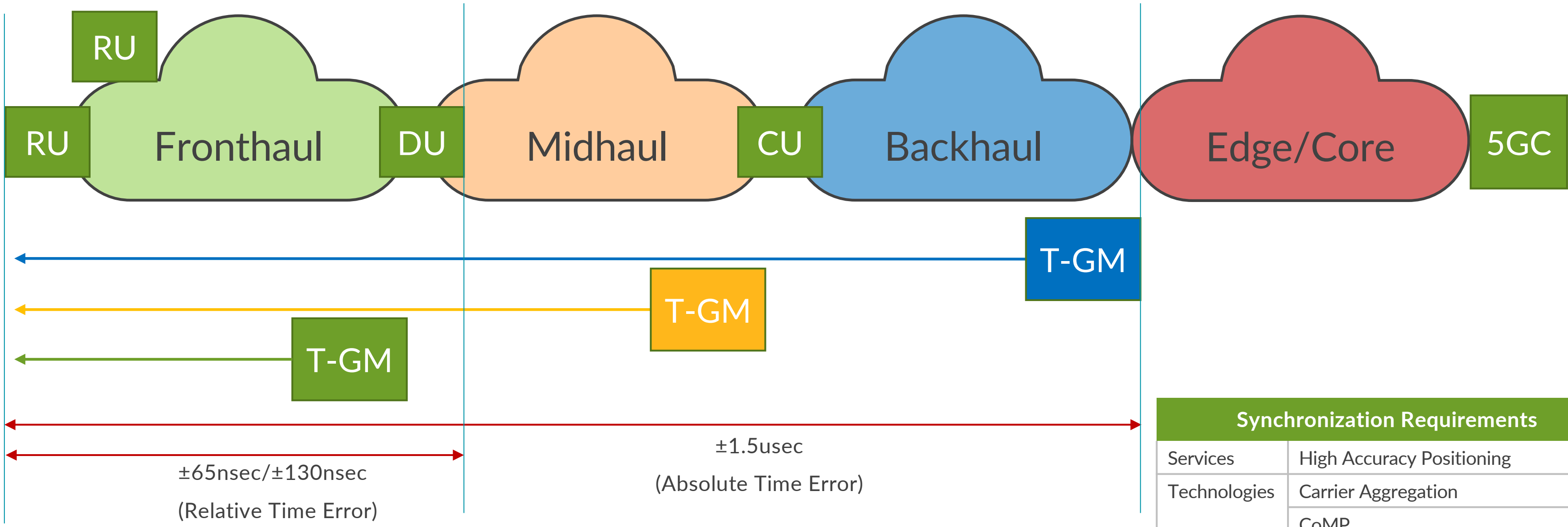


- Data centres transition from private to public cloud infrastructure
- More Intelligent applications moving to network edge. (Ex: Internet of Things, online gaming, 5G connectivity and autonomous driving)
- A single autonomous car produces **4–5 TB/day**.
- Deploy energy-efficient data centres with smaller carbon footprints
  - According to the International Energy Agency, data centres worldwide consumed about 200 terawatt-hours (TWh) of electricity in 2020, nearly 1% of global electricity demand.
- Training jobs are highly sensitive to latency and packet loss. 10usec delay can cause 20% connectivity issues and training loss.

5G Timing Requirements

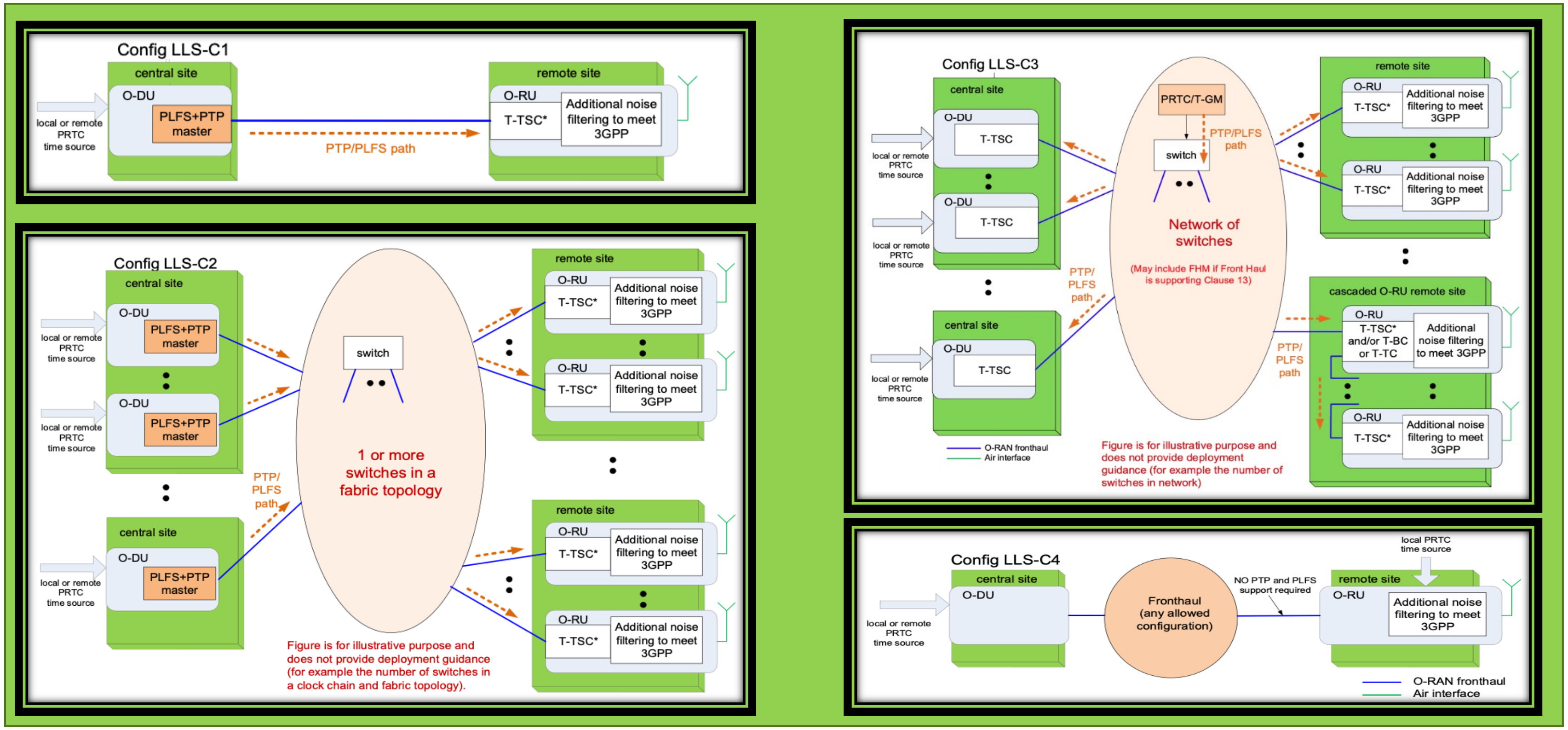
Technology/Application	Phase accuracy	Relative/Absolute
Intra-band contiguous Carrier Aggregation (CA)	+/- 65 ns	Relative
Intra-band non-contiguous CA	+/- 130 ns	Relative
Inter-band CA	+/- 130 ns	Relative
Coordinated multi-point (CoMP) with Joint Transmission (JT)	+/- 130 ns	Relative
High accuracy positioning service (All RRU/ AAU connected to same DU)	10 ns	Relative
Self-driving/Autonomous cars	< 5 ns	Relative
Transmit diversity Category A+	+/- 32 ns	Relative
MIMO (category A+)	+/- 32 ns	Relative
eCPRI (IEEE 802.1CM)	+/- 130 ns	Relative

5G Overview



Synchronization Requirements	
Services	High Accuracy Positioning
Technologies	Carrier Aggregation
	CoMP
	5G Frame Structure
Architecture	Backhaul, Midhaul and Fronthaul

O-RAN Deployment Configurations



Source: O-RAN-WG4.CUS.0-v08.00 specification

O-RAN Timing Network Limit

LLS-C2 Configuration

Network Limit at O-RU UNI	Expected TE Limit	
	Enhanced RU	Regular RU
Packet max|TE_L|	≤ 1465ns	≤ 1420ns
Packet max|TE_{RL}|	≤ 60ns (FR2)	≤ 100ns
	≤ 190ns (FR1)	
1PPS max|TE_L|	≤ 1465ns	≤ 1420ns
1PPS max|TE_{RL}|	≤ 60ns (FR2)	≤ 100ns
	≤ 190ns (FR1)	
Frequency Limit (For O-DU Class-A)	≤ 36ppb	≤ 36ppb
Frequency Limit (For O-DU Class-B)	≤ 32ppb	≤ 32ppb

LLS-C3 Configuration

Network Limit at O-RU UNI	Expected TE Limit	
	Enhanced RU	Regular RU
Packet max|TE_L|	≤ 1100ns	≤ 1100ns
Packet max|TE_{RL}|	≤ 60ns (FR2)	≤ 100ns
	≤ 190ns (FR1)	
1PPS max|TE_L|	≤ 1100ns	≤ 1100ns
1PPS max|TE_{RL}|	≤ 60ns (FR2)	≤ 100ns
	≤ 190ns (FR1)	
Frequency Limit (For O-DU Class-A)	≤ 36ppb	≤ 36ppb
Frequency Limit (For O-DU Class-B)	≤ 32ppb	≤ 32ppb

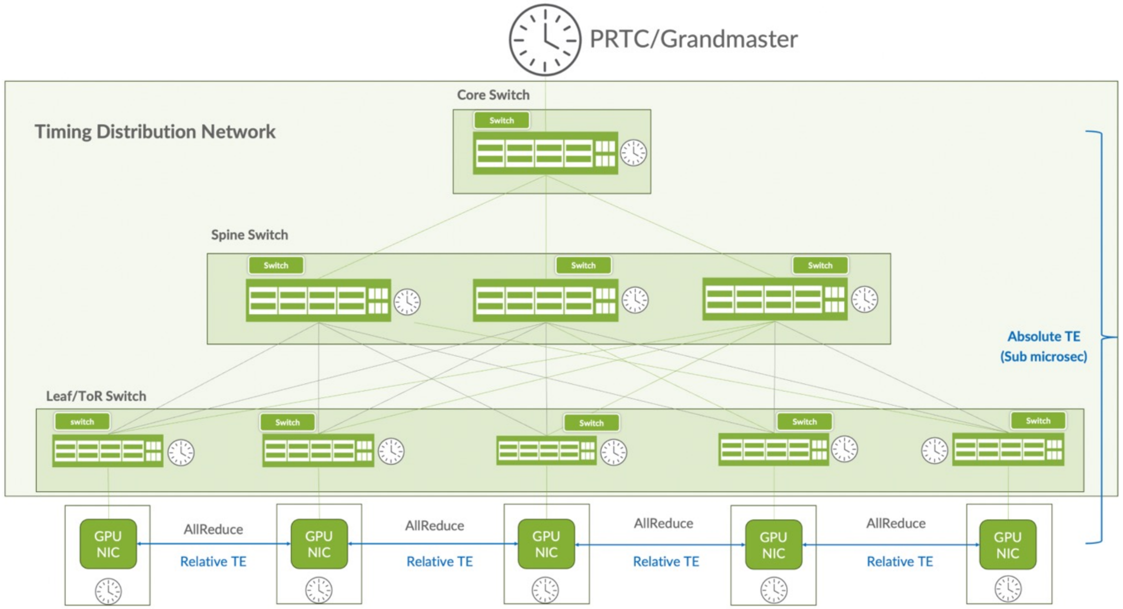
AI for Networking vs Networking for AI

- ❑ Using AI/ML models to make the network smarter, autonomous, self-optimizing
  - ✓ 5G RAN, Core, Transport, Operators
- ❑ Building network infrastructure that supports AI workloads, AI training clusters, Fabrics.
  - ✓ AI datacenters, MEC AI clusters

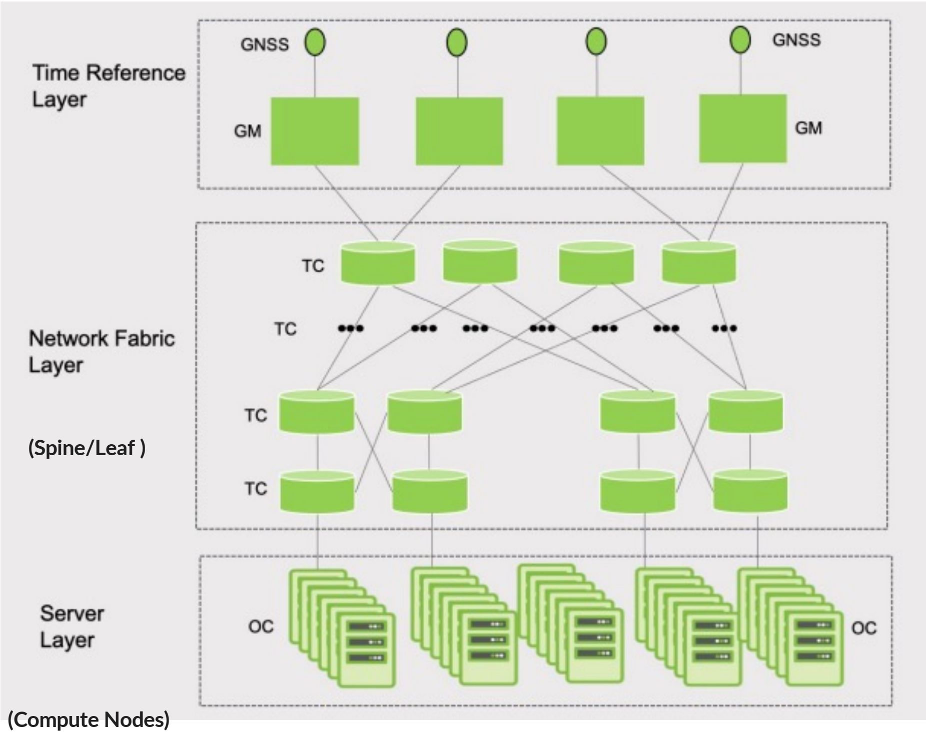
Use cases	
RAN Optimization ✓ Beamforming ✓ PRB scheduling ✓ Dynamic TDD ✓ Mobility/Handover Prediction	AI Training Custers over 5G
Network Optimization ✓ Congestion prediction ✓ Path selection ✓ OWD based congestion control ✓ Drift prediction ✓ Time Error Prediction	MEC AI inference

Timing Distribution in AI Cluster

- ❑ AI clusters interconnect thousands of GPUs and servers- **high-bandwidth, low-latency fabric** built from Ethernet, InfiniBand, or RoCEv2 switches.
- ❑ **Fabric node**, forwards compute data (model parameters, gradients) and **timing information** (PTP, SyncE, or TSN time sync).
- ❑ CLOS or Dragonfly interconnect with spine-leaf layers **(BC or TC)**, ensuring deterministic multi-path routing.



TAP-Model1



(Compute Nodes)

Source: OCP Profile for DC (Model1)

Time Error Requirements	Accuracy
The maximum absolute time error between any two OCs	≤5 μs
The maximum absolute time error between a GM and any OCs	≤2.5 μs
The maximum time error between any 2 GMs	≤100 ns
The maximum time error generated by a TC	≤100 ns

Source: OCP Profile for DC (Model1)

Typical Applications	Relative Time error requirements between TSCs
Distributed databases, applications profiling	≤5 μs
High-Frequency Telemetry, Multi-node performance analysis tools	≤2.5 μs
Congestion control based on one way delay, Time synchronized collective communication	200 ns

Source: ITU-T G-Sup.DCSync

ITU-T Profiles consideration and IEEE1588.1

PTPoE link local multicast **derived from** G.8275.1 profile
PTPoIP unicast based on IEEE P1588.1 – **Client Server PTP (CSPTP)** with full timing support from the network utilizing transparent clocks.

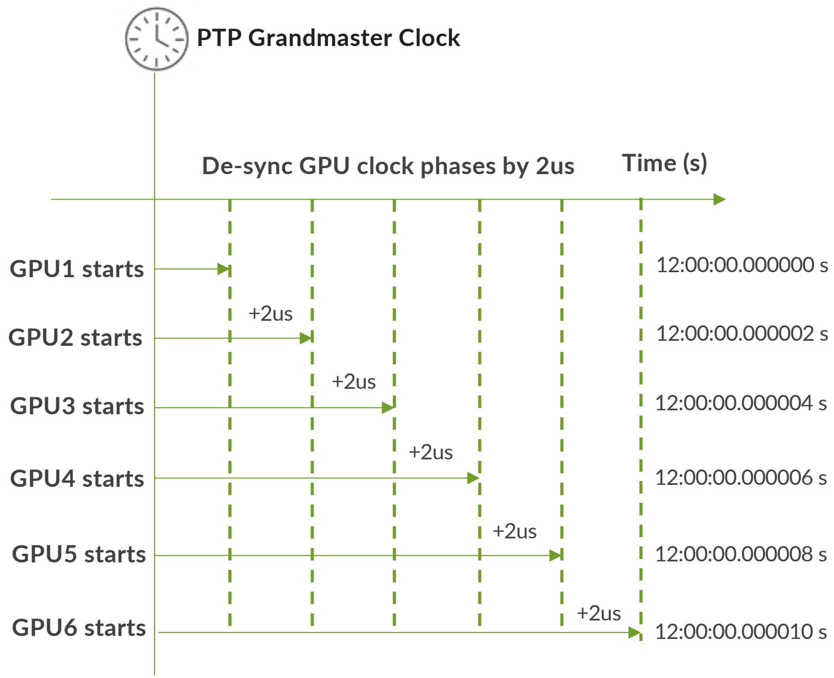
Time aware collective communication

Deterministic collective communication

- AI training clusters involve **tens of thousands of GPUs** exchanging gradients every few microseconds.
- Each GPU independently computes the loss and local gradients for its data slice.
- GPUs must synchronize their updates through a collective communication operation called AllReduce.
- Precision timing can ensure that these exchanges happen with sub-microsecond accuracy.

Sync-to-desync Power Control

- ❑Stagger the GPUs start at precise time, avoiding high current burst.
- ❑Dynamically schedule loads to machine based on power consumption.
- ❑Dynamically schedule loads to machine based on the computing efficiency.



One way Delay

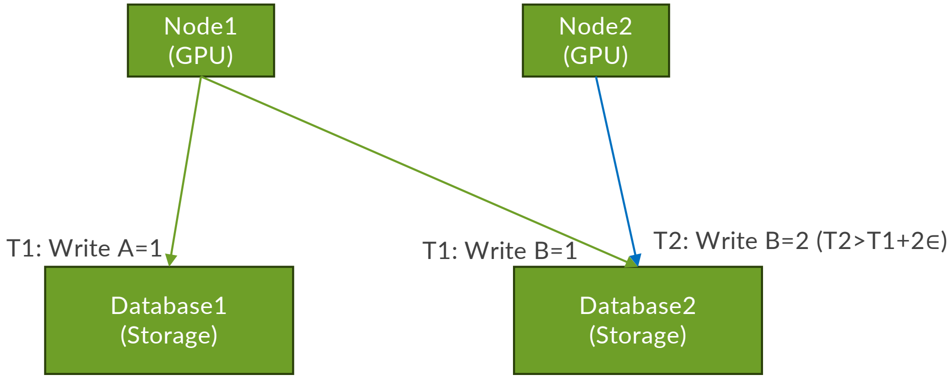
With sub-microsecond-level clock differences across devices, we can measure one-way delay, locate packet losses, and identify per-hop latency bursts.

Enables synchronized network snapshots.

Better congestion signal to delay-based congestion control to differentiate between forward and reverse path congestion

Distributed Transactional Databases

Time Uncertainty Bound (∈)



- T1 (Node 1) is a distributed transaction that writes A=1 on Database 1 and B=1 on Database 2.
- T2 (Node 2) later wants to write B=2 on Database 2.
- '∈' is the **max clock error** (time-uncertainty) relative to real time.
- To be **certain** T2 happens after T1 in *real time*

- T2 > T1 + 2∈
- Better clock sync → **smaller ε** → **smaller commit-waits** and shorter safe gaps
- Drive ∈ from, say, **100 ns** with PTP/DC-grade sync, that **2ε safety margin shrinks to 200 ns**, materially cutting write latency.

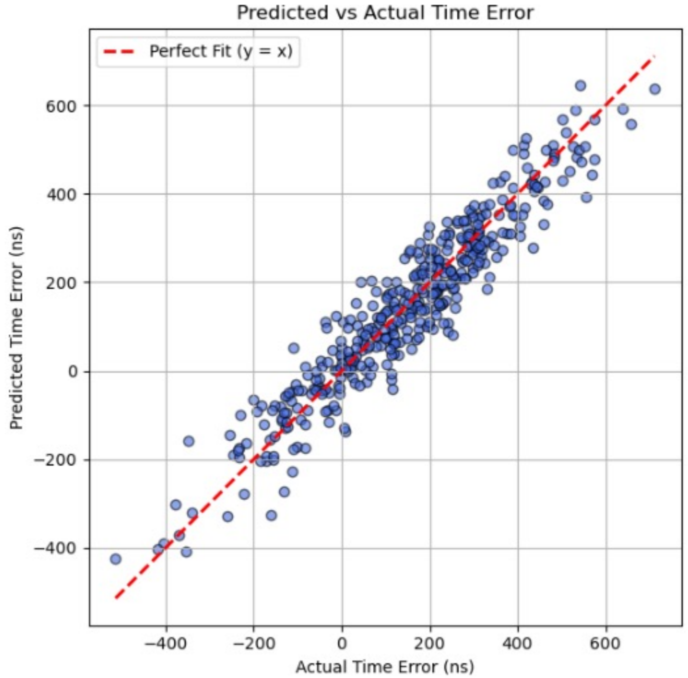
FaRMv2, an RDMA-based transactional system, observes the median transaction delay can drop by 25% if we improve ∈ from ~20µs to 100ns.
CockroachDB can significantly reduce the retry rate when ε drops from 1ms to 100ns.

Predictive Modelling in Sync using ML and DL

Predictive Modelling: Learn from the data sets and predicting.
Adaptive Systems: Compensate for errors due to dynamic changes.
Optimized Control: Optimized servo loops to minimize clock drift and deviation.

Model Performance

R-squared	Adj R-squared
0.899598	0.896751



R-squared/Adjusted R-squared

- The model explains **~90 % of total variance** in time-error measurements.
- Only ~10 % remains unmodeled (random PDV, jitter spikes, or unmeasured features).
- Adjusted $R^2 \approx 0.897$. This indicates high explanatory power even after accounting for feature count.

Model Quality

Very good, captures most deterministic contributors to TE.

Questions


THANK YOU
JUNIPER NETWORKS
Driven by Experience™