Synchronising the cloud to support AI applications

Presenter: Ullas Kumar

Outline

- Background
- AI and HPC applications
- Convergence of networks
- Clocking architectures
- Deployment options

Traditional applications

Managing databases

- ~microsecond precision to reduce the Commit-wait cycles of nodes

Optimising search functions

- Sub-microsecond time synchronization, for cache invalidation -> consistency

Real-time content delivery

- Precise time stamping to assist the content updates, user interactions, and data changes
  -> Enables Event Ordering

Faster responses to end users

- Time synchronisation reduces the variance in response times

AI workloads need synchronised time

Synchronising parallel workloads

- AI workloads, especially training, involve synchronised parallel jobs across nodes. - 33% of AI elapsed time is spent waiting for the network (Meta)

Time and order-sensitive workloads

- Frame-accurate correlation of video data across distributed encoders/decoders
- Maintaining Audio/Video Sync across processing streams
- Coordinating parallel video processing pipelines

Real-Time processing

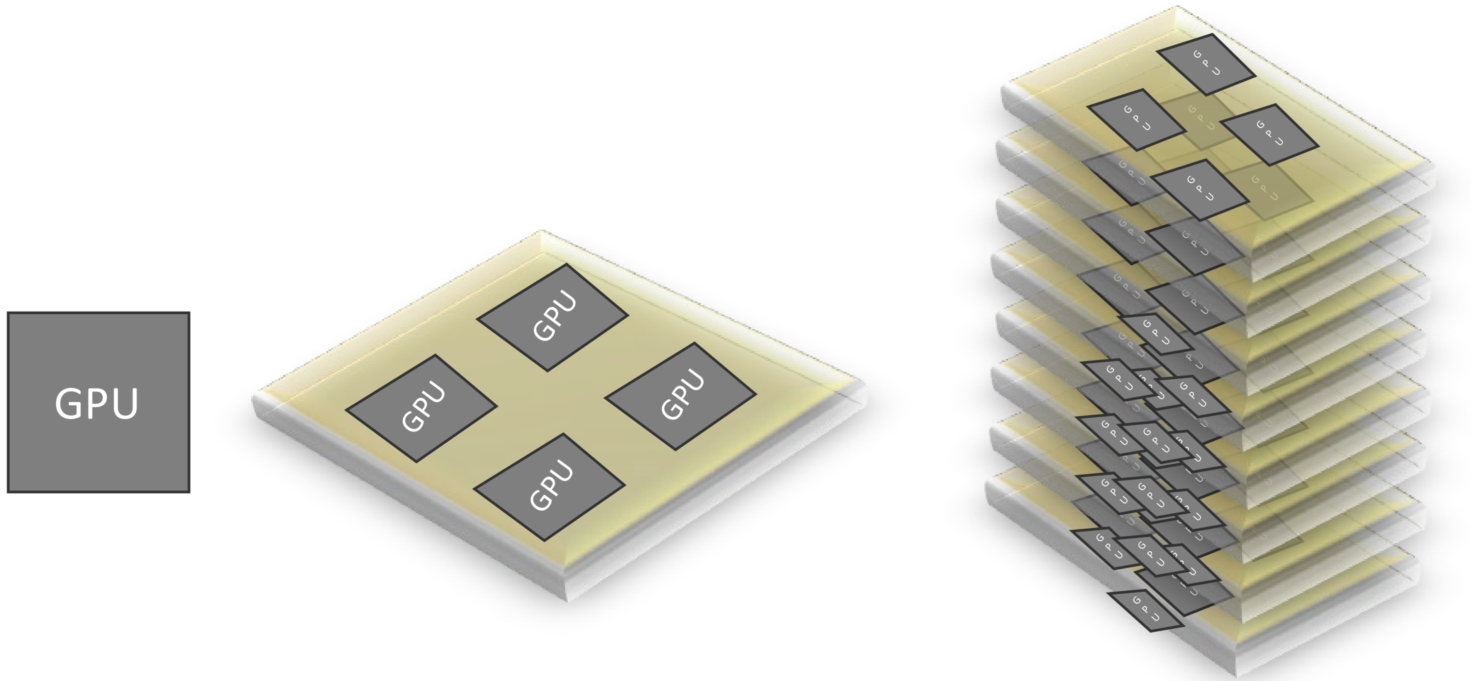
- Surveillance systems, autonomous vehicles, or live video analytics

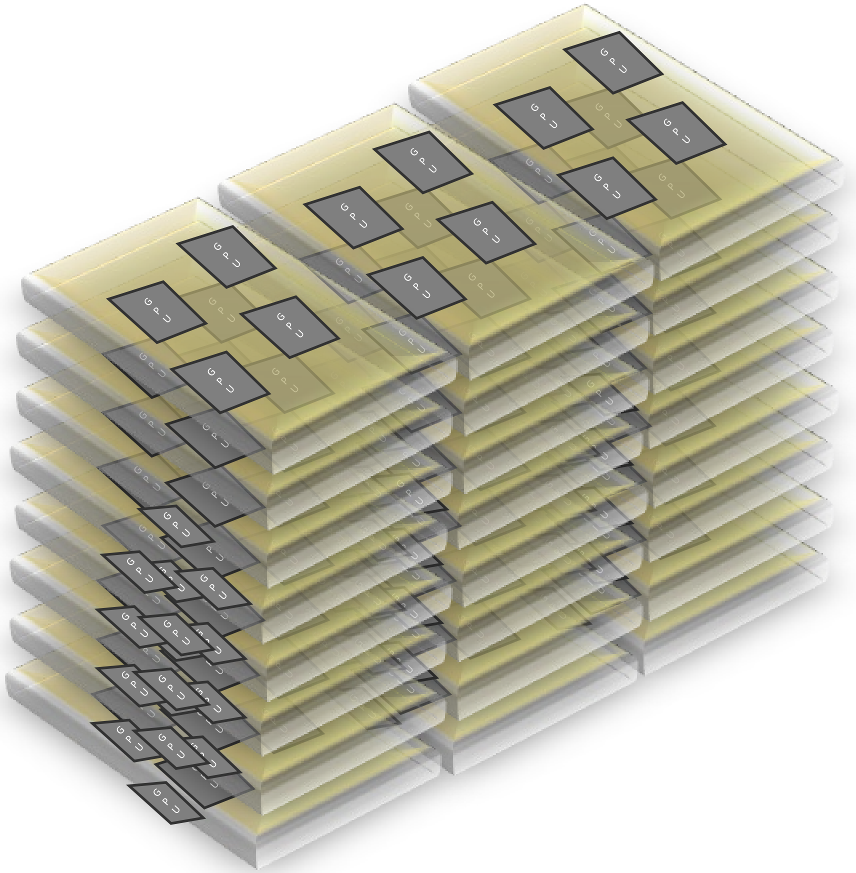
Enhanced User experience for 3D simulation

- Precise video timing across multiple nodes

Overview of AI hardware architectures

Scaling up as one machine – AI clusters of GPUs.



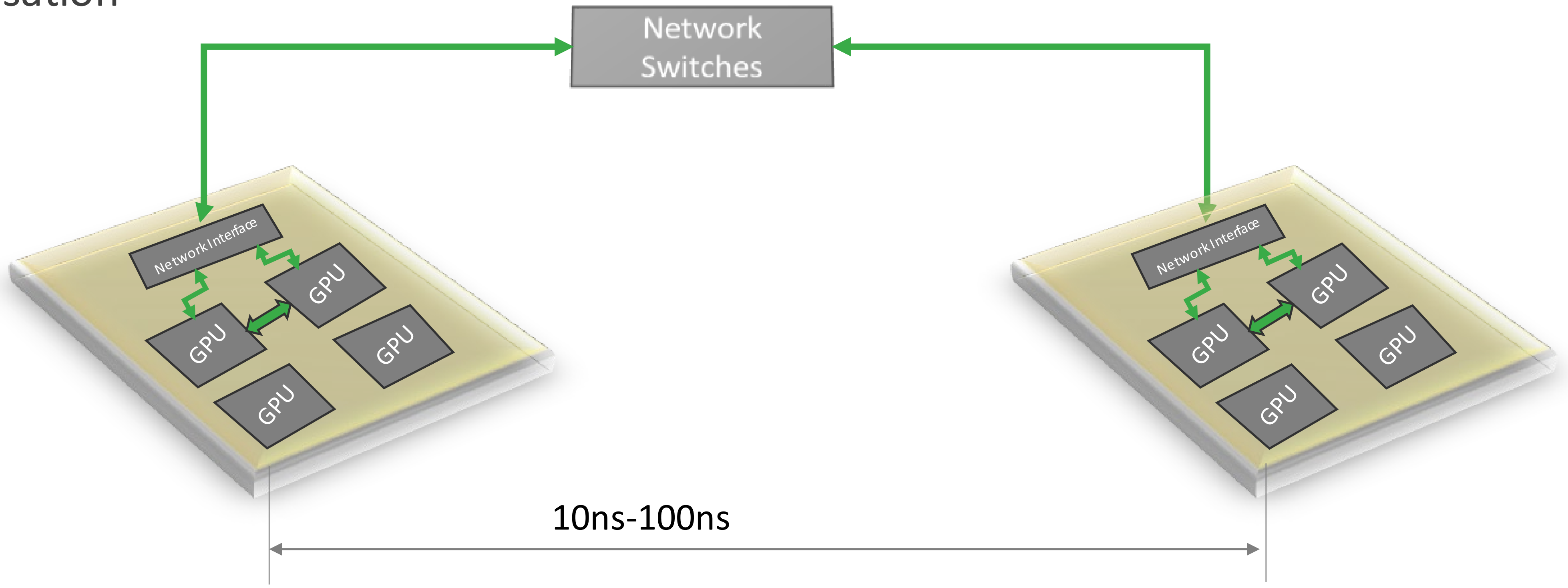


Needs a single time domain

SoCs **Trays** **Racks** **Pods**

Synchronisation overview

GPU interconnects, Network Interfaces and Switching elements to support high-performance synchronisation



Distribution Unit (DU) fits Data Centre (DC)

Matching capabilities

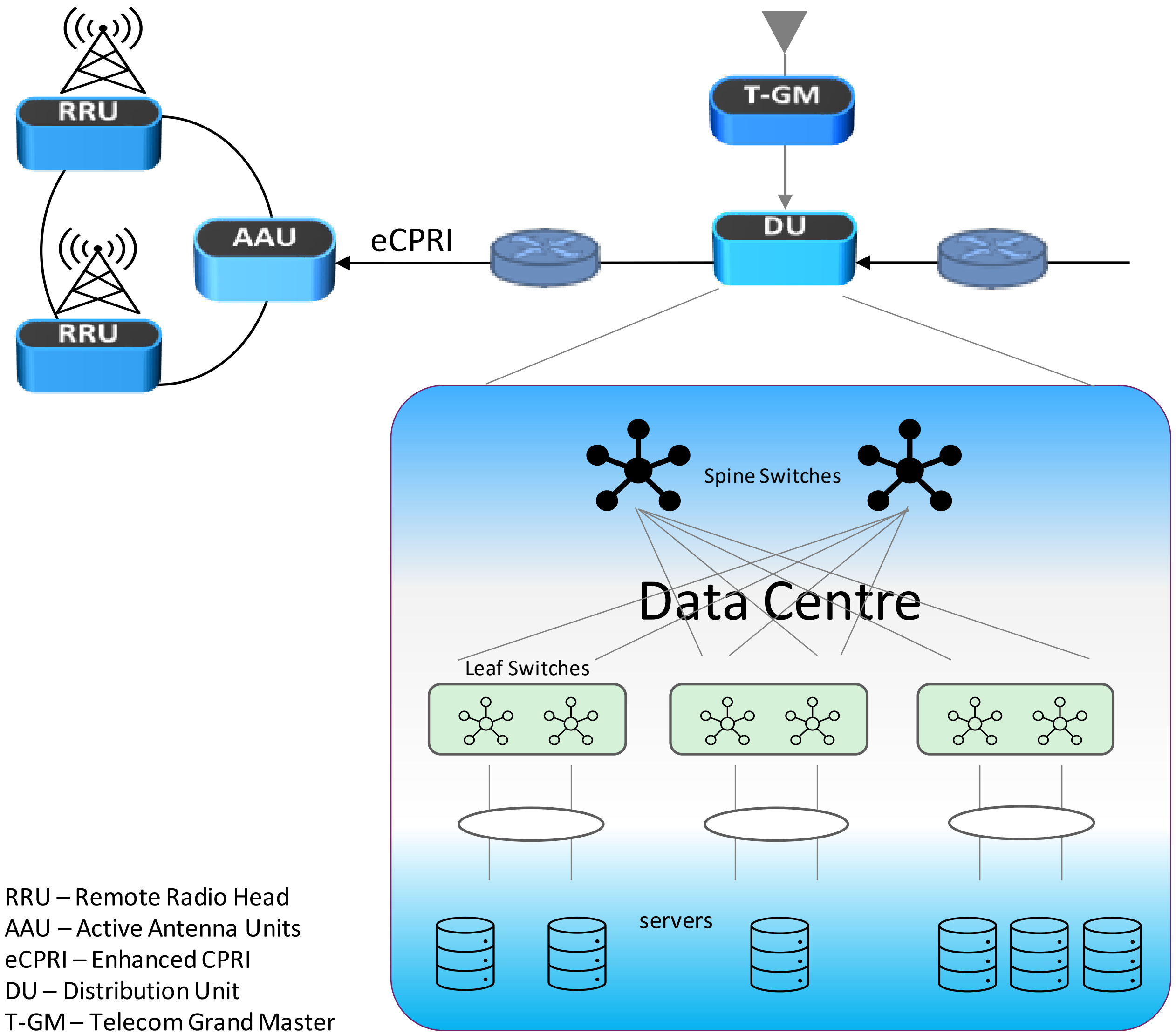
- Computing, storage and networking

Proximity to end applications

- URLLC use cases, autonomous driving
- Real-time application off-loading

Private 5G advantages

- Secure Industrial applications with secure on-prem infrastructure
- Integrated AI-ML applications



RRU – Remote Radio Head
AAU – Active Antenna Units
eCPRI – Enhanced CPRI
DU – Distribution Unit
T-GM – Telecom Grand Master

Data centre architectures

Various architectures

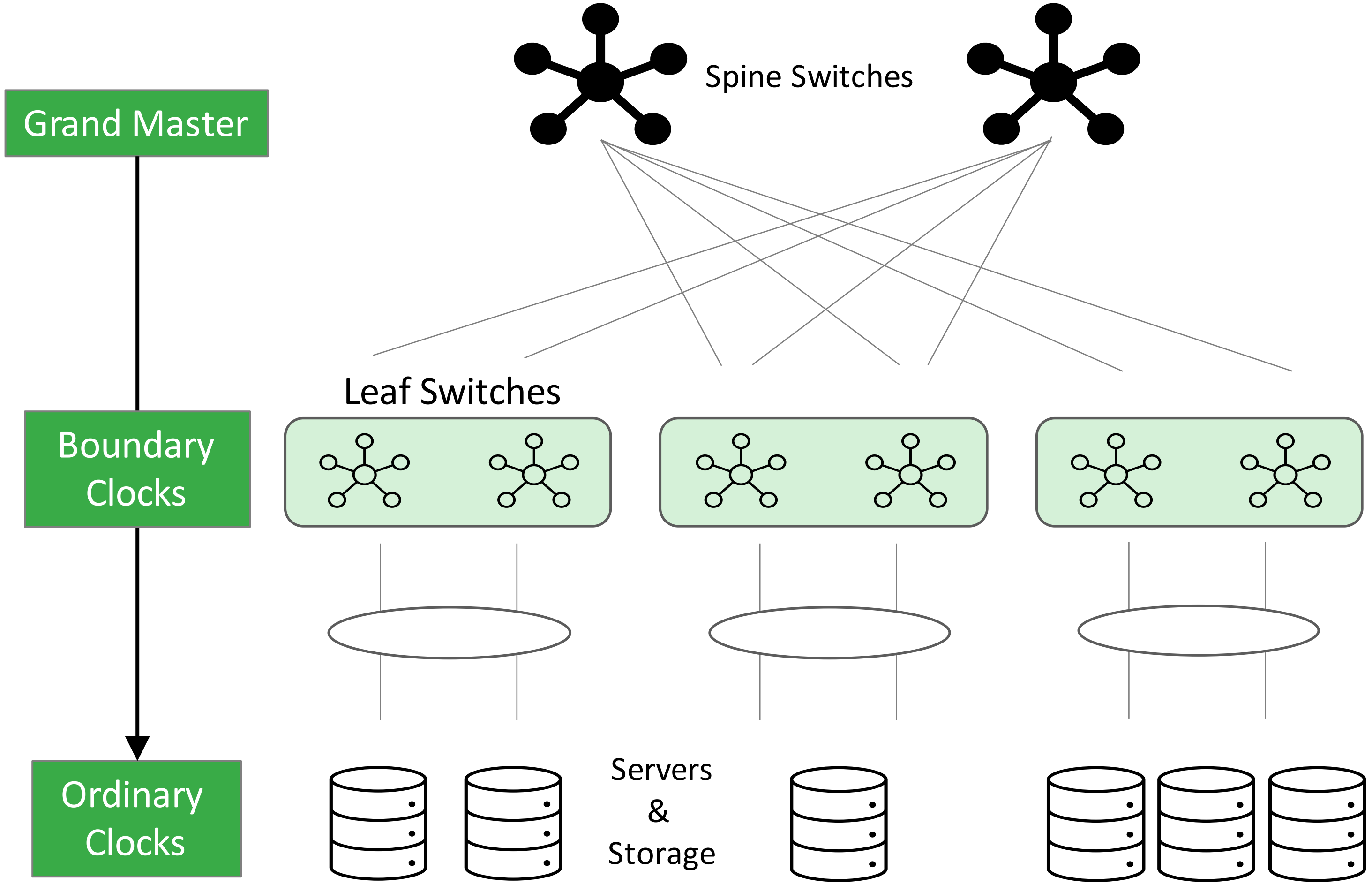
- Spine and Leaf architecture is popular

Synchronisation exists in data centres

- Managing databases
- Optimising search functions
- Real-time content delivery
- Faster responses to end users

Synchronisation performance

- In millisecond range
- Not enough to meet telecom requirements



Telecom DC synchronisation

Objective:

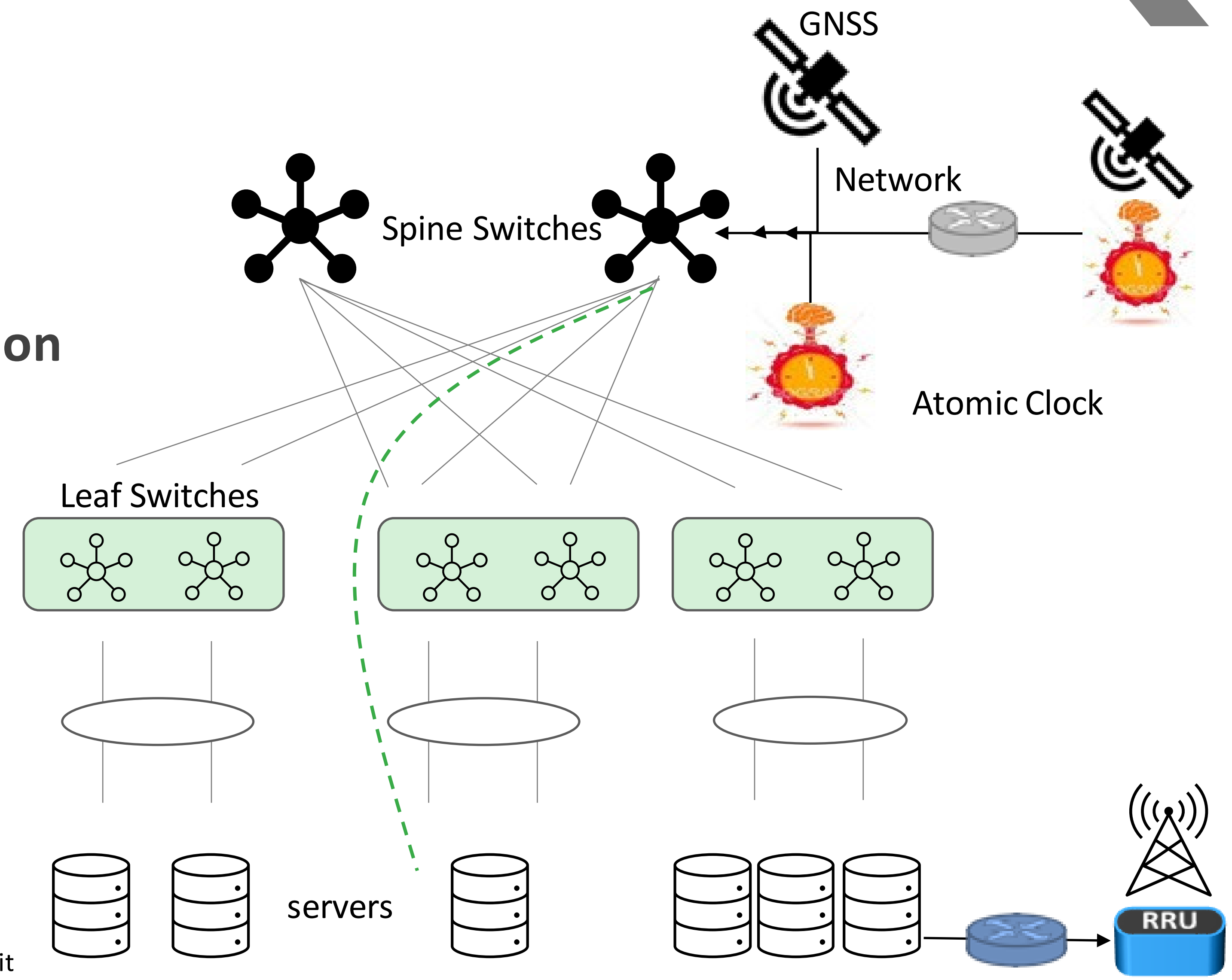
- Use the DC as DU/CU systems in 5G
- Drive Radios from Data Centres

Enable telecom level synchronisation

- Atomic, GNSS or Network Clocks

Method to transfer the clock

- Within data centres

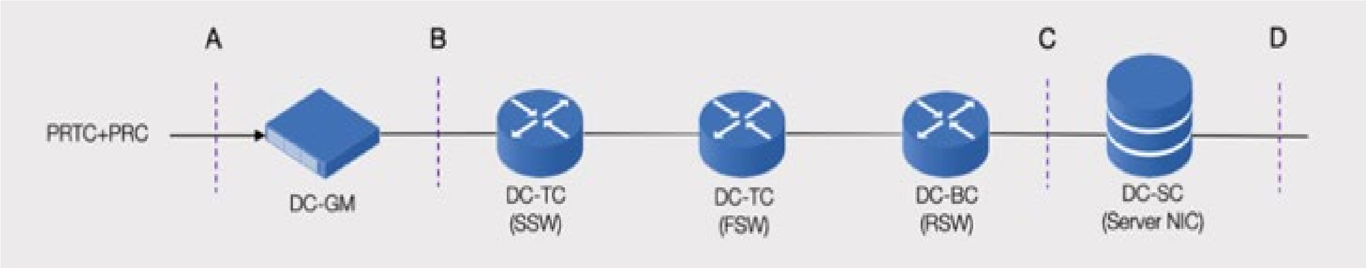


CU – Central Unit
DU – Distribution Unit
DC – Data Center
RRU – Remote Radio Unit

Clocking schemes

Transparent clock architectures

- Most common systems within the data centre use transparent clock techniques



Reference Point	Limit
A	Max [TE] ≤ 100 ns
B	Max [TE] ≤ 250 ns
C	Max [TE] ≤ 1000 ns
D	Max [TE] ≤ 5000 ns

GM – Grand Master
FSW - Fabric Switches
SSW - Spine Switches
RSW - Rack Switches
BC – Boundary Clock
SC – Slave clock
NIC – Network Interface Card

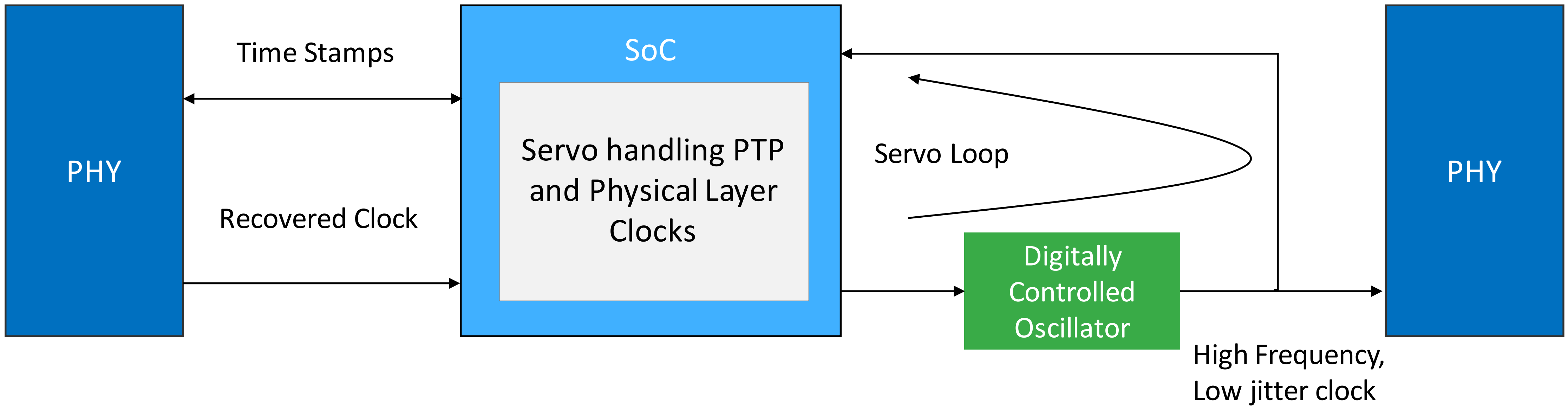
HRM by TAP

- Recommends the use of transparent clocking architecture

Synchronising physical layer

Synchronising the physical layer gives access to traceable clock

- Every single node is synchronised



Summary

AI data centres and HPC systems are moving towards high-precision clocking

- Sub-microsecond precision enables a variety of applications

Implementations of boundary clocks have enabled higher performance than those of transparent clocks.

High-stability, low-cost digital control oscillator solutions enable such implementations.

Q & A

www.rakon.com