Time Appliance Project

Vision / Challenges/ Target / Roadmap

Consume. Collaborate. Contribute.

OPEN Compute Project

PTP Challenges in Datacom

Dotan D. Levi

- with Ahmad Byagowi, Michel Ouellette and Georgi Chalakov

Aug 12, 2020

Outline

- Why?
- What is different/Unique?
- Challenges
- Goals
- Q&A

Next Generation Data center – Requires Time Sync

- Services that can be boosted by Time synchronization
  - Message service: simplify the “At-most one message” guarantee
  - Authentication Tickets: today in minutes, with Time sync service - down to milliseconds
  - Cache Consistency: Precise clocks can synchronize the invalidation w/o Communication…
  - External consistency makes the service indistinguishable from a single machine
    - Google spanner, “DC like a PC” up to x80 performance
  - Commit window: distributed DB consistency is accelerated by time sync.
    - Higher availability, and decoupled from the network latency
- DC applications spread to the Edge
  - Edge require time sync
  - Edge consider part of the service …
- More…

Time Service in Data center

Trade Off Pyramid

Scalability

Time Service in Data center

Trade Off Pyramid

DCs are constantly growing.
Any designs must scale to hyperscale.
The PTP clock tree and performance must remain at the same quality and should not get affected by the up scaling

Scalability

Time Service in Data center

Trade Off Pyramid

DCs are constantly growing.
Any designs must scale to hyperscale.
The PTP clock tree and performance must remain at the same quality and should not get affected by the up scaling

Also includes Virtualization, Overlay Network, Security, and Hybrid Deployment with previous services such as NTP.

Scalability

Time Service in Data center

Trade Off Pyramid

DCs are characterized by the application, running in scale on common infrastructure.
It is acceptable to have an outage, or accuracy loss. However, a mechanism should **inform** the application of such service degradation.
In case of timing and synchronization, the service monitoring should be available even for a single network transaction.

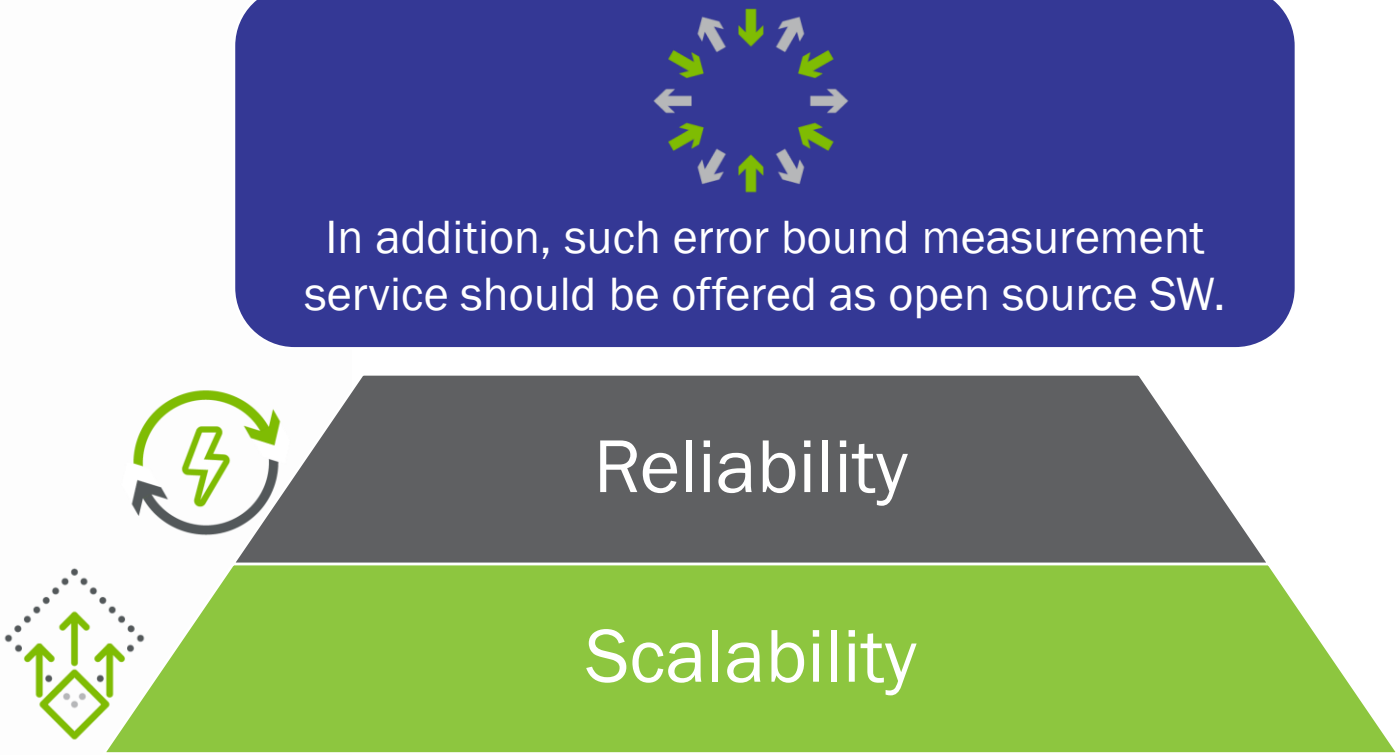
Today, there is not a well-known or standardized method known for real time error bound measurement, (not to mention, in-scale).

Time Service in Data center

Trade Off Pyramid

DCs are characterized by the application, running in scale on common infrastructure.
It is acceptable to have an outage, or accuracy loss. However, a mechanism should **inform** the application of such service degradation.
In case of timing and synchronization, the service monitoring should be available even for a single network transaction.

Today, there is not a well-known or standardized method known for real time error bound measurement, (not to mention, in-scale).

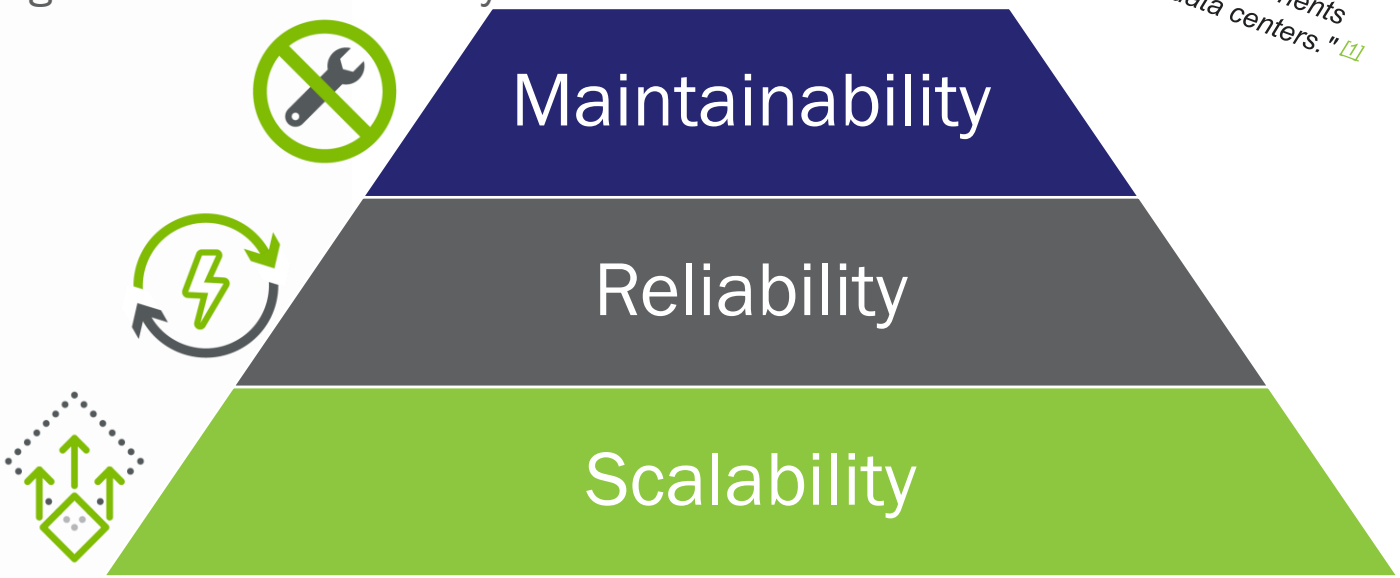


Time Service in Data center

Trade Off Pyramid

In a hyperscale DC, PODs are pulled out for maintenance. The operation team is not aware on the application restrictions. PTP service, like any other services, must maintain a self auto discovery and allow unaware maintenance.
The service itself should be maintained by a high level of observability.

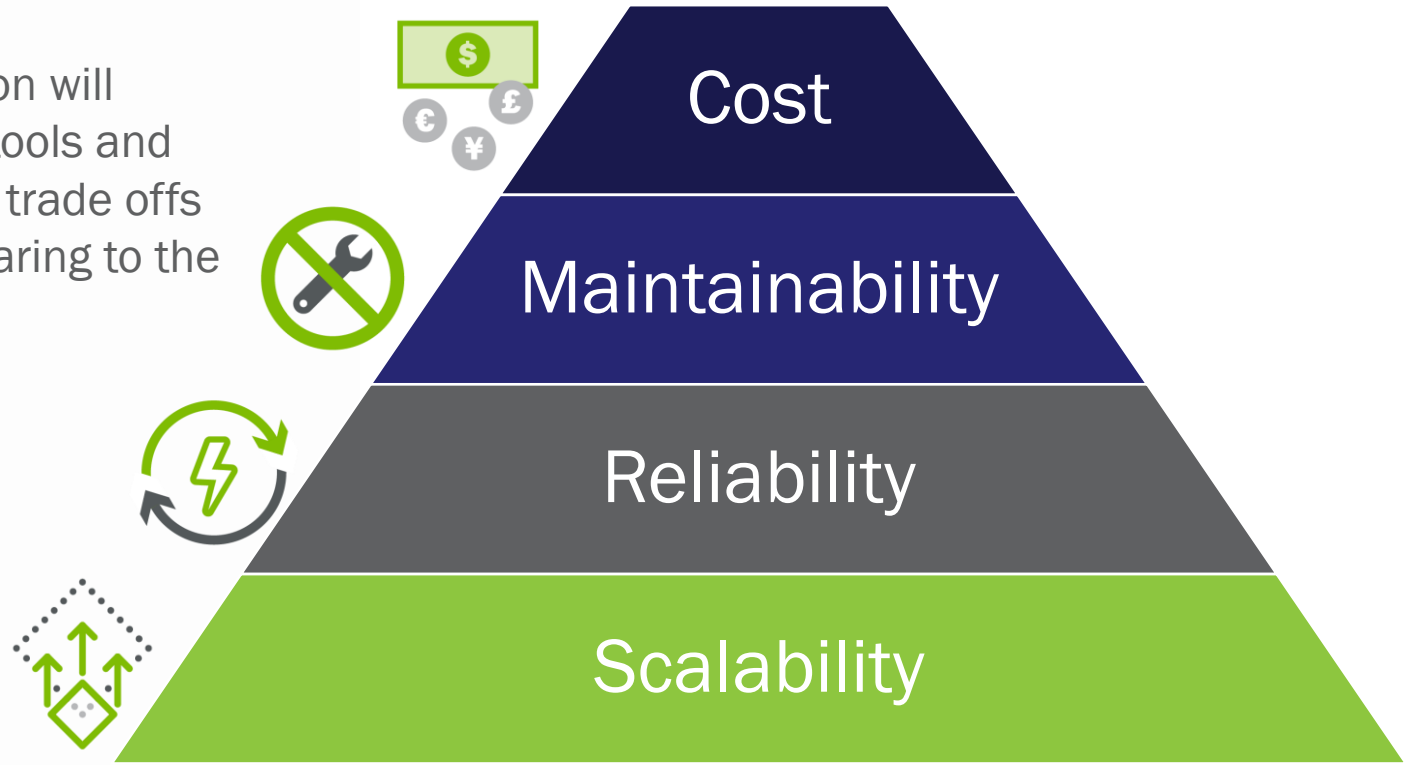
*A **point of delivery**, or **PoD**, is "a module of network, compute, storage, and application components that work together to deliver networking services. The PoD is a repeatable design pattern, and its components maximize the modularity, scalability, and manageability of data centers." [1]*



Time Service in Data center

Trade Off Pyramid

Due to the hyperscale nature of DCs, the cost factor is the next important thing.
To name some, Oscillator Selection will determine CAPEX, however sync tools and monitoring will save OPEX. Those trade offs are different in a Datacom Comparing to the Telecom

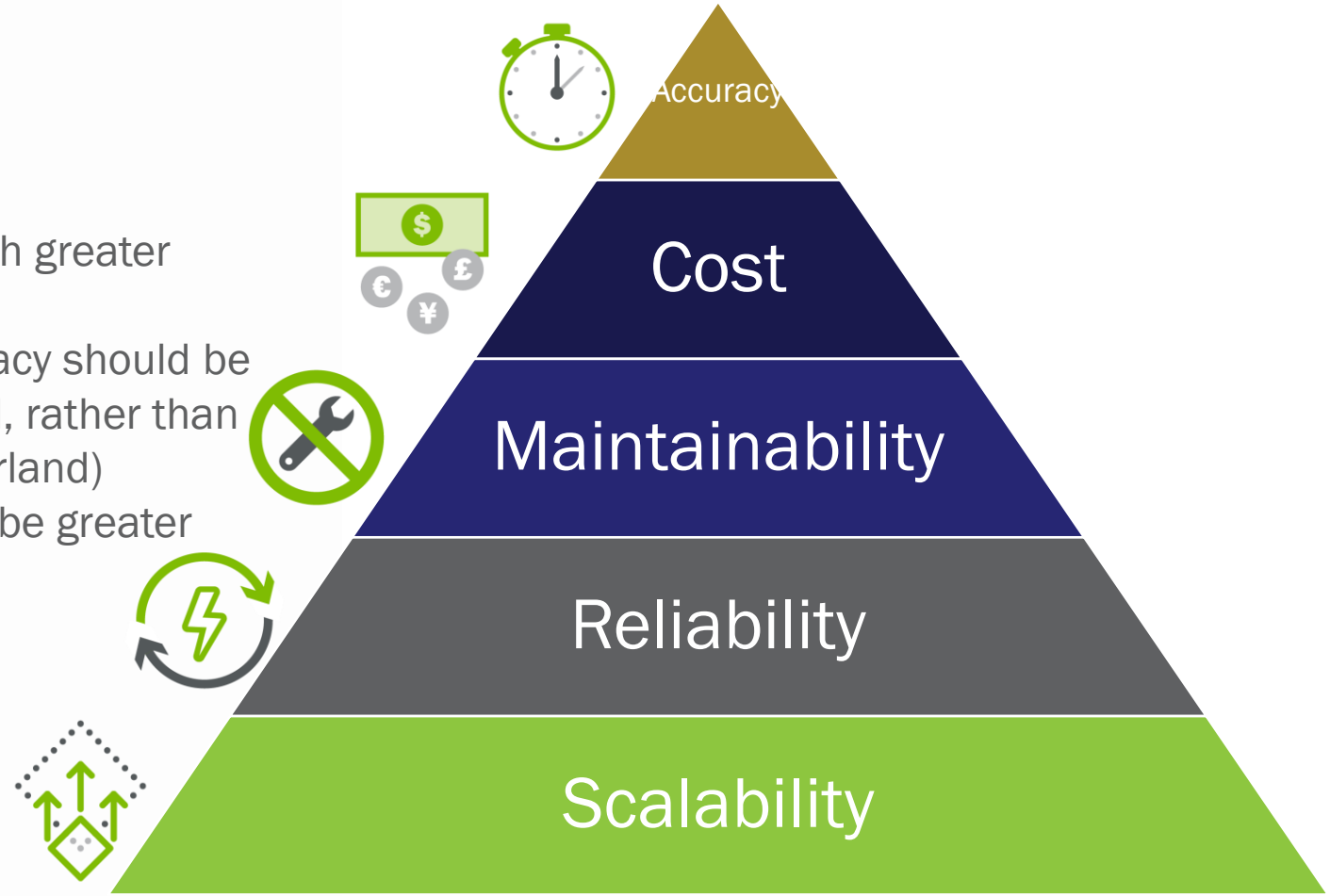


Time Service in Data center

Trade Off Pyramid

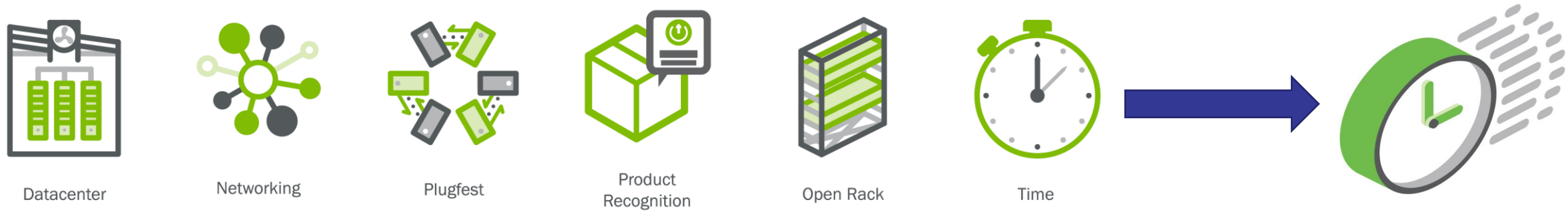
Finally, comes the accuracy.

Todays PTP technologies allow much greater accuracy than needed in DCs.
However, we believe that the accuracy should be defined **also** in the application level, rather than in the individual clock. (PHC vs userland)
Application Error bound will always be greater than the clock Error bound itself.



TAPs Vision:

A design/blueprint to allow fast, reliable, scalable, and maintainable deployment of PTP in the DC, at the desired cost/performance option.

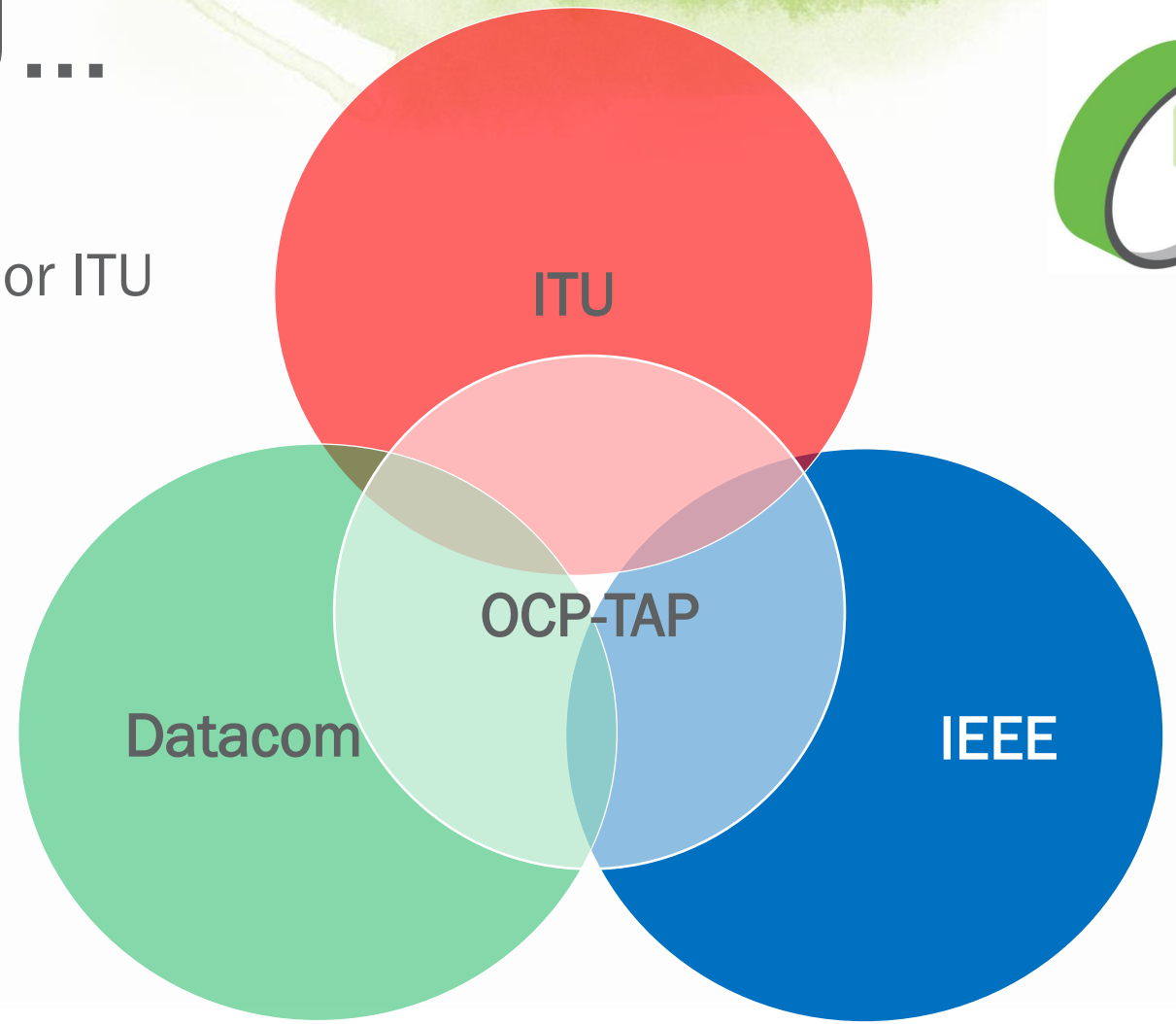


What do we intend to Do?

And what we don't

TAP & IEEE, ITU…

- We do not intend to replace IEEE nor ITU
- Rather, rely-on and customize



Topics that require definition

- DC-PTP Profile
  - Clock Tree
  - Noise Filters
  - Measurement Methods
  - Testing Methods
  - Testing Suites
- Open source Grandmaster and architecture
- Open source implementation
- Oscillator classes
- OCP Timing reference designs

A Sneak peek into clock tree issues

- Wide clock tree
- Even a single BC scale is challenging
  - In an OCP Multi-Host each network link contain multiple nodes
  - Switches with 256 ports-> 1000s of ordinary clocks
- Longer Clock chains and Noise propagation
  - A single noise event will exist in the system for a long time
  - Probability of noise event during previous noise effects is high!
    - Can cause infinite instability
- Reliability
  - Cannot rely on a single GM
    - Can be in a maintenance window
    - Wish to avoid operational and maintenance awareness

THANK YOU

Consume. Collaborate. Contribute.

OPEN
Compute Project