中国移动
China Mobile

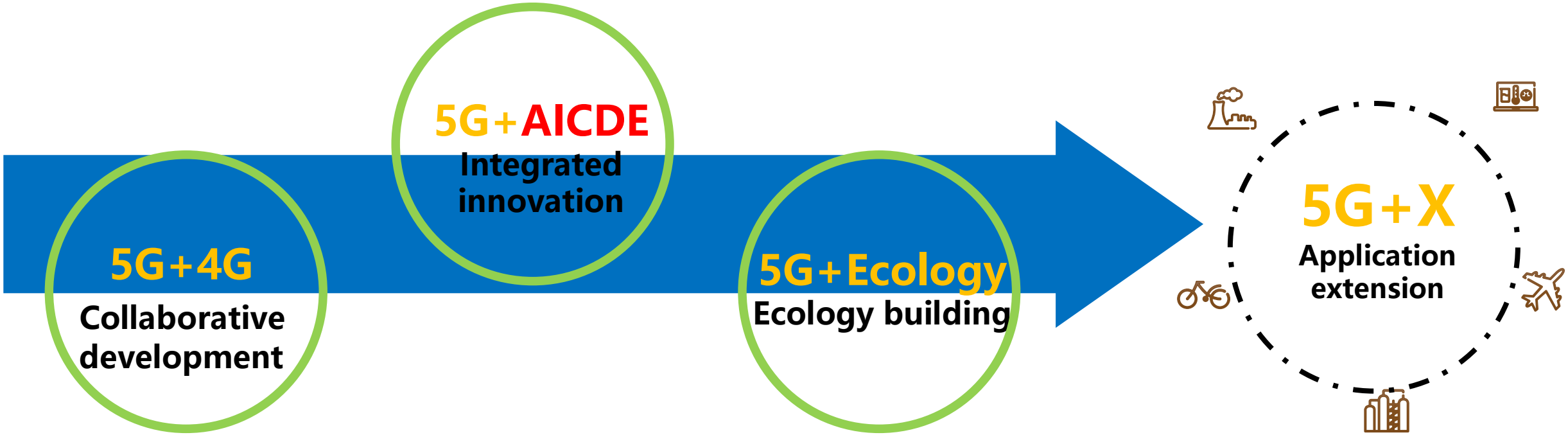
New Requirements and Challenges Raised by Edge Computing for Future Network

Dr. Liang GENG
Manager of Edge Computing Research, China Mobile Research Institute
Chair of Architecture WG, Network 5.0 Alliance
Oct,2019, Geneva

- Background of edge computing
- Requirements for future network evolution
- Solution: High Precision Computing Network

Edge computing is the core for”5G+”plan of China Mobile

Edge computing realizes the extension of connection to computing power from center to user premises

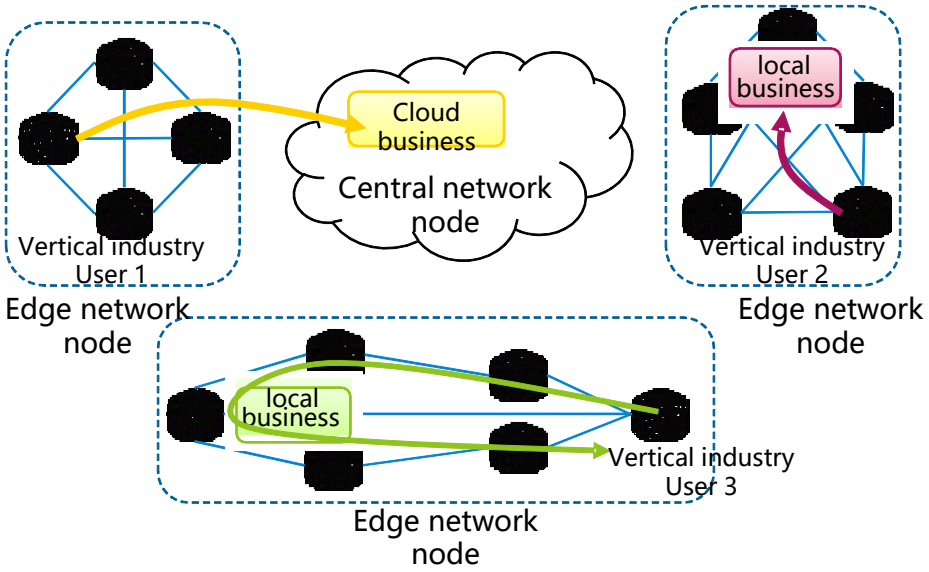


Key to success : ICT integrated Infrastructure + Open Platform + Vertical Application

Demand for Edge Computing of Industrial IoT

Localized data processing and time critical application build up strong requirement for Edge Computing

Decentralized business/traffic model



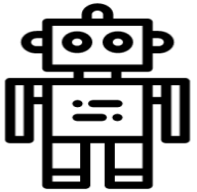
- On- Premise application (Security concerns)
- Massive volume of traffic caused by HD video sensors (Traffic Offloading)

Optimal latency and jitter

Industrial control

Telemedicine

Robot

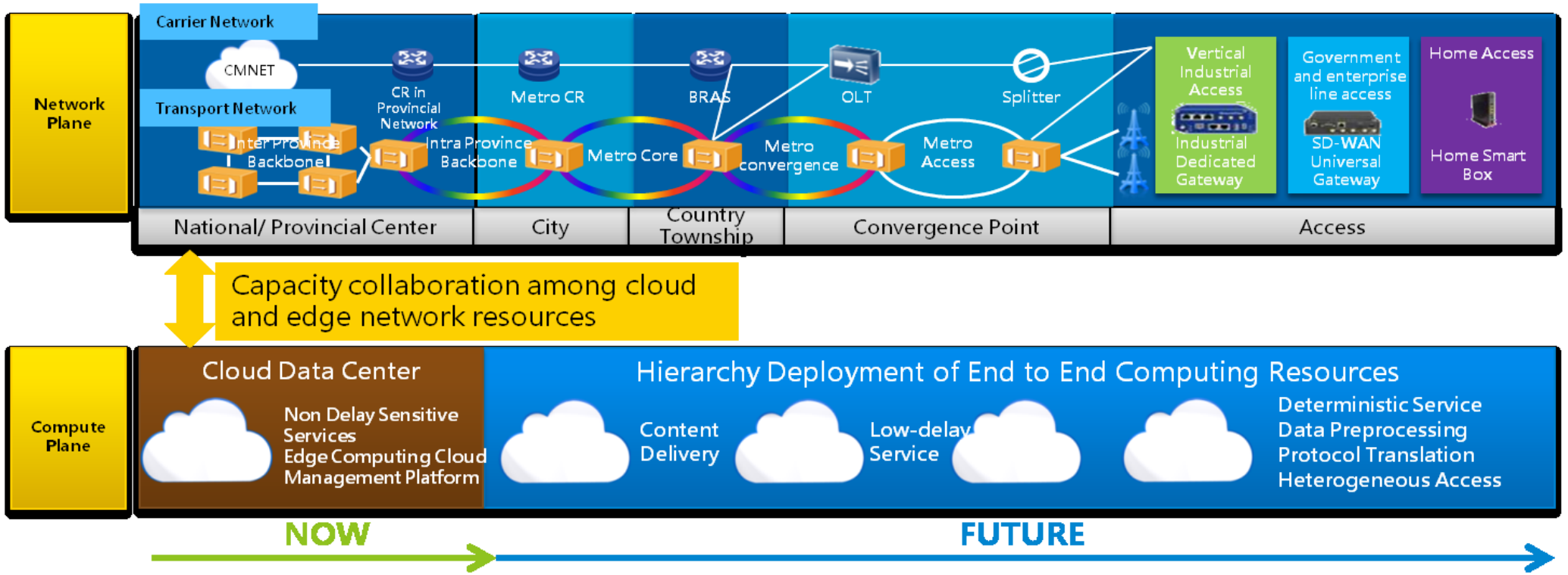


- **Businesses with strict delay requirements** are processed near client is more beneficial to improve the user experience
- **Data pre-processing requirements,** such as filtering, encapsulation, etc., reduce the bandwidth requirements of edge network to center network
- **Edge AI** requires more computing power at the edge

4 Key Motivation for Edge Computing Deployment
Security, Traffic off-loading, Latency and High efficient computing

The Evolution of ICT Infrastructure Caused by Edge Computing

Individual Network and Computing infrastructure base are evolving to a converged and intertwine plane



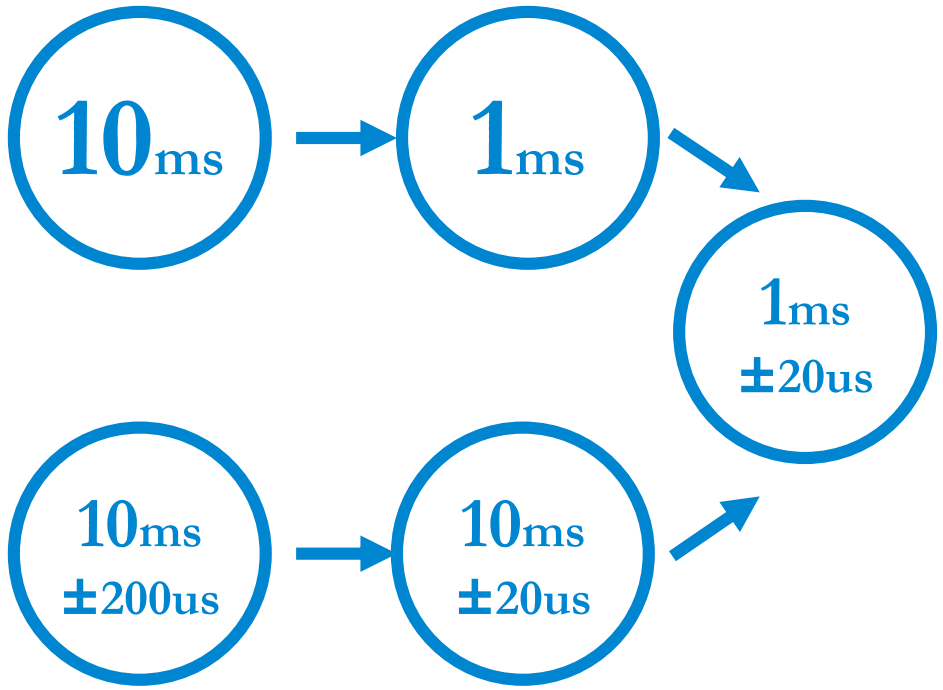
Edge computing promotes the integration of development of network plane and computing plane，and builds an industry-wide, fully connected, intelligent infrastructure

- Background of edge computing
- Requirements for future network evolution
- Solution: High Precision Computing Network

Requirement 1： High-precision connectivity

Best Effort to ***Low Latency*** to ***Precise Latency and Jitter***

Low latency is the optimization of the absolute value of time delay



target: Low delay+ high certainty

Precise/Deterministic latency is the optimization of standard deviation

Industrial control **Remote surgery** **Robot**

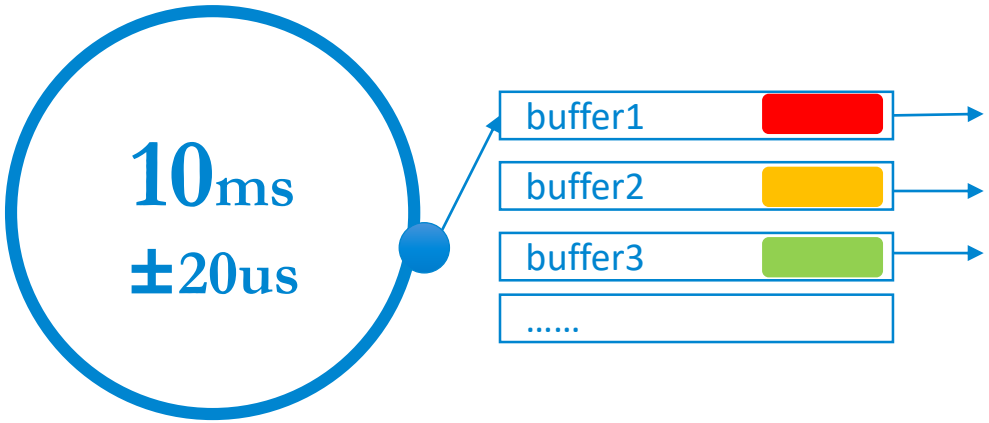
- Industrial control, telemedicine, robot and other scenarios need end-to-end precise signaling transmission
- Deterministic latency is the **basis** of the cooperation of multiple control systems, such as multi-arm linkage, flexible manufacturing, human-machine interaction, etc.
- Deterministic latency puts forward a new **measurement dimension** for carrying network technology, **from "in-time" to "on-time"** .

Requirement 1： High-precision connectivity

Precision should be further evolve to multi-demensional

On-time transfer

Dedicated time slot+ periodical scheduling



Now - Deterministic latency

deterministic jitter

deterministic delay

deterministic bandwidth

deterministic loss rate

deterministic connection

deterministic path

Deterministic…

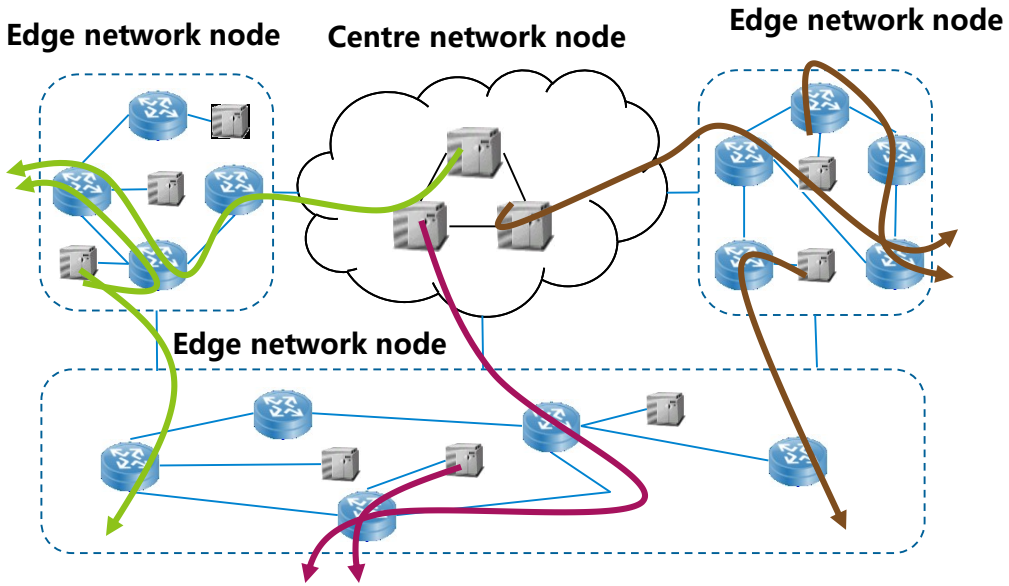
Future – high precision network

High-precision network can achieve certainty in multiple dimensions, and can achieve comprehensive SLA guarantee with explicit boundaries

Requirement 2: Widely distributed and heterogeneous computing

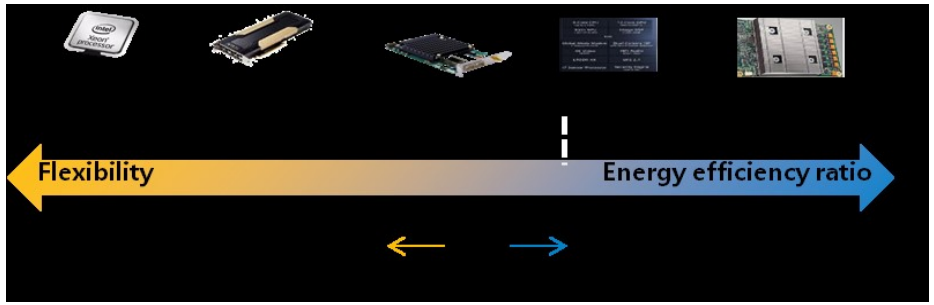
With the development of 5G and industrial Internet, massive connectivity is driving traffic decentralization and requires widely distributed and heterogeneous computing.

Distribution of Computing Resources



Heterogeneity of Computing Resources

With the development of artificial intelligence, there is higher requirements for computing capability of network. The computing power evolves from general to heterogeneous, to achieve the best computing efficiency and user experience.



Ubiquitous and heterogeneous computing envision a further integrated Computing + Network infrastructure

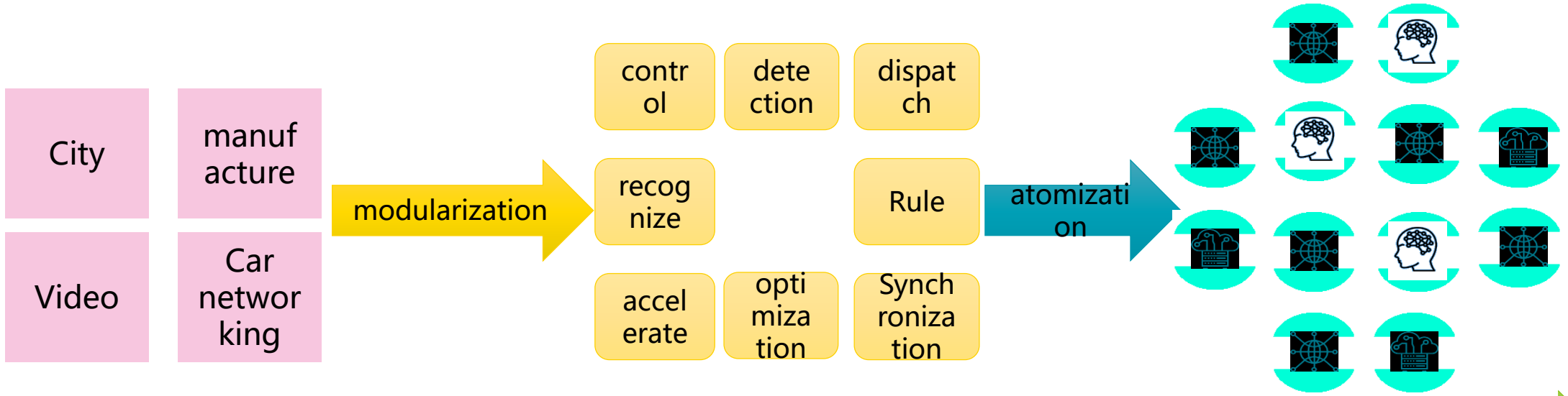
Requirement 2: Widely distributed and heterogeneous computing

Ubiquitous computing helps realize function as a service (FaaS) and forwarding as processing.

Past – Walled-garden Apps

Now – Common APIs (Middle PaaS)

Future – Fully service-based and atomic functions



Toward Function as a service

Integration from infrastructure level to Service/Algorithm/Function Level

Requirement 3: Edge-edge & edge-cloud coordination

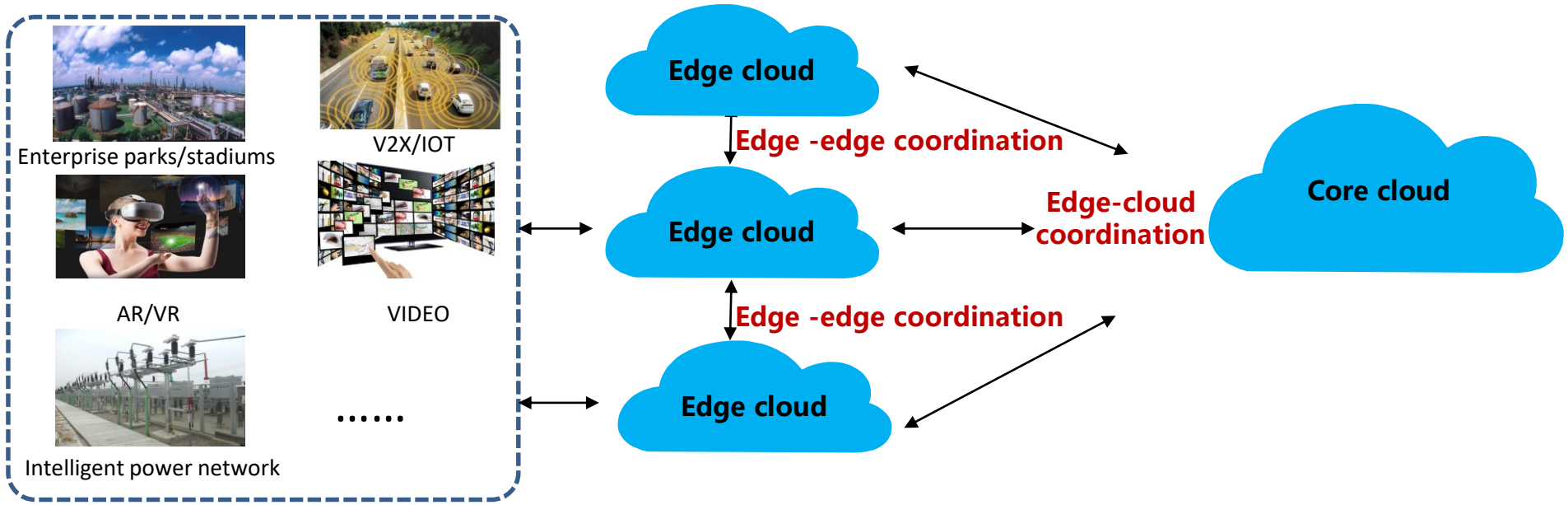
The widely distributed computing sites needs to be coordinated by a more intelligent network. This coordination requires computing-network joint optimization

Edge-edge coordination

For interactive and delay-sensitive applications such as AR, VR, with edge-edge coordination, the task could be separated into micro-tasks and scheduled to multiple edge sites to guarantee the delay requirements.

Edge-cloud coordination

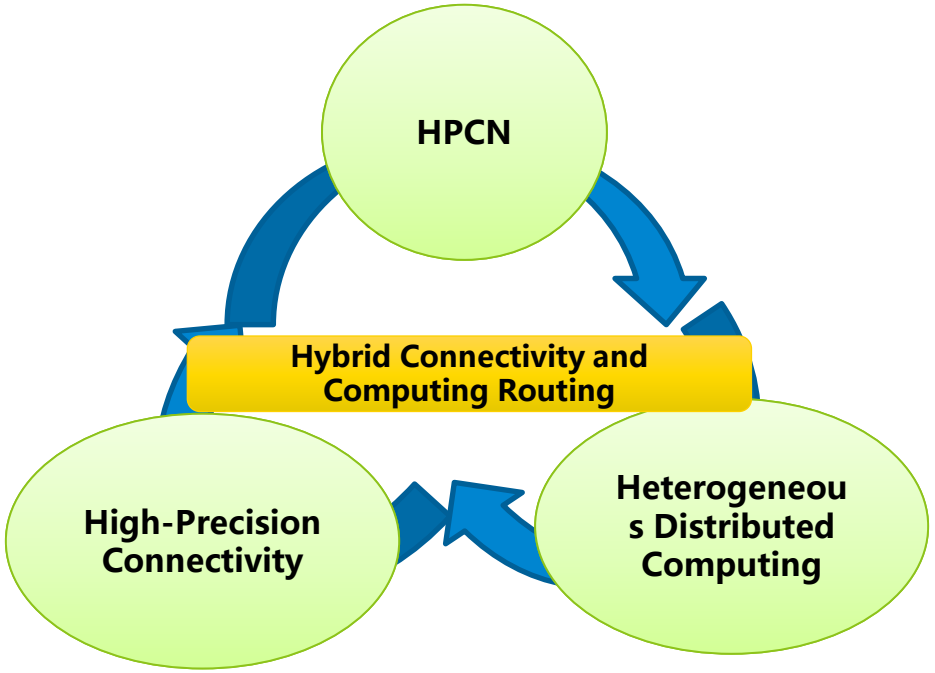
There are many AI applications such as face recognition requiring low latency and industrial AI scenarios requiring data locally processing, etc, require model inference locally at edge site and model training at cloud to achieve the best user experience and network performance.



- Background of edge computing
- Requirements for future network evolution
- Solution-High Precision Computing Network

High Precision Computing Network (HPCN)

A New Architecture – Converged network and computing with High Precision Connectivity and Ubiquitous Computing

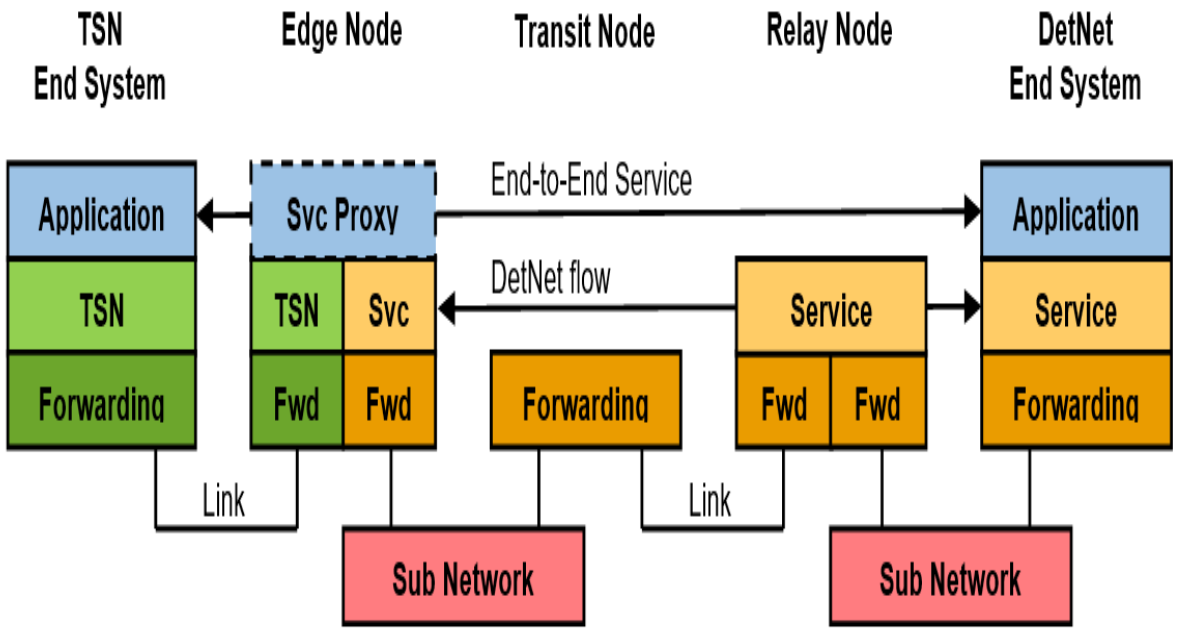


- High-Precision Connectivity (HPC)：Optimized latency and jitter， together with multi-domain determinism including throughput, security, routing path, bandwidth, etc.
- Heterogeneous Distributed Computing (HDC): provides widely distributed computing resource that are interconnected by HPC.
- Hybrid Connectivity and Computing Routing (HCCR): Computing information is advertised the network to enable joint scheduling and optimization. Traffic can be steered to preferred edge computing site according to performance requirement and HPC and HDC status

HPCN set the base architecture for fully converged and distributed ICT system in the future

High Precision Connectivity (An Detnet Example)

Reserve necessary resources for deterministic traffics and **Forwarding plane-scalable deterministic forwarding mechanism** enable bounded latency in DetNet.

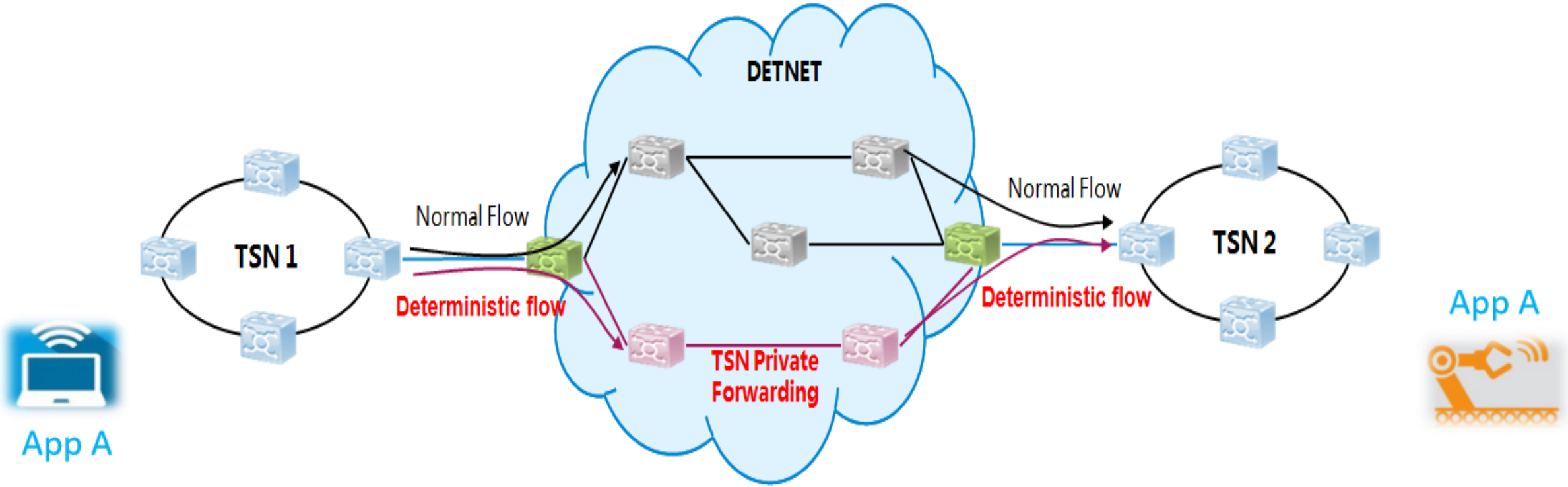


DetNet Enabled Network

- DetNet (enabled) end systems: "host" (IETF), and an "end station" (IEEE 802). Sources or destinations of DetNet flows.
- DetNet relay nodes: includes a DetNet service sub-layer function and DetNet forwarding sub-layer functions as well.
- DetNet edge nodes: is a DetNet relay node that acts as a source and/or destination at the DetNet service sub-layer. (It is analogous to a Label Edge Router (LER) or a Provider Edge (PE) router.)
- DetNet transit nodes: it operates at the DetNet forwarding sub-layer, provides congestion protection over those paths. (An MPLS LSR is an example of a DetNet transit node.)

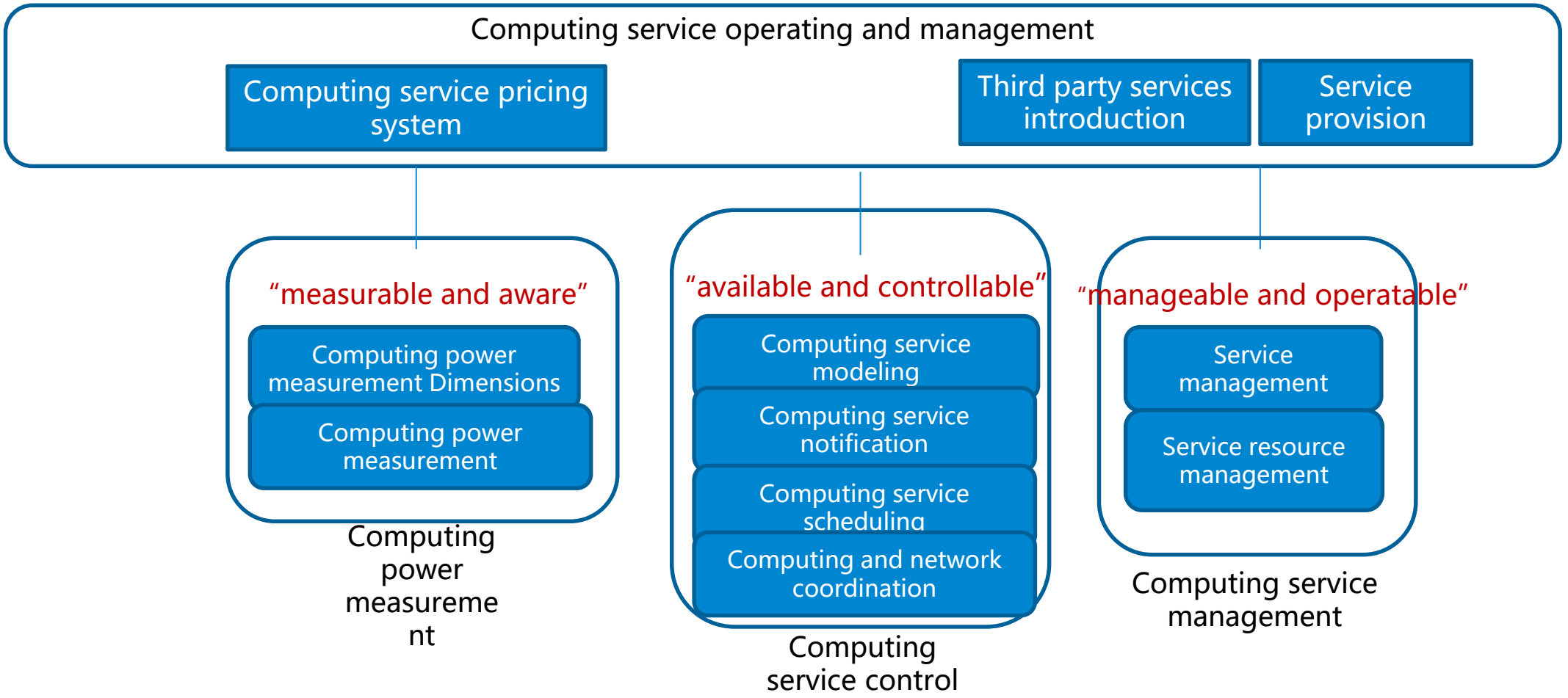
High Precision Connectivity (An Detnet Example)

Implementing interconnection to TSN islands realizes deterministic service over large-scale network.



Scalability of DetNet in large-scale networks including stateless of traffic forwarding, no need for precise time synchronization between devices is beneficial for DetNet applied in the industry.

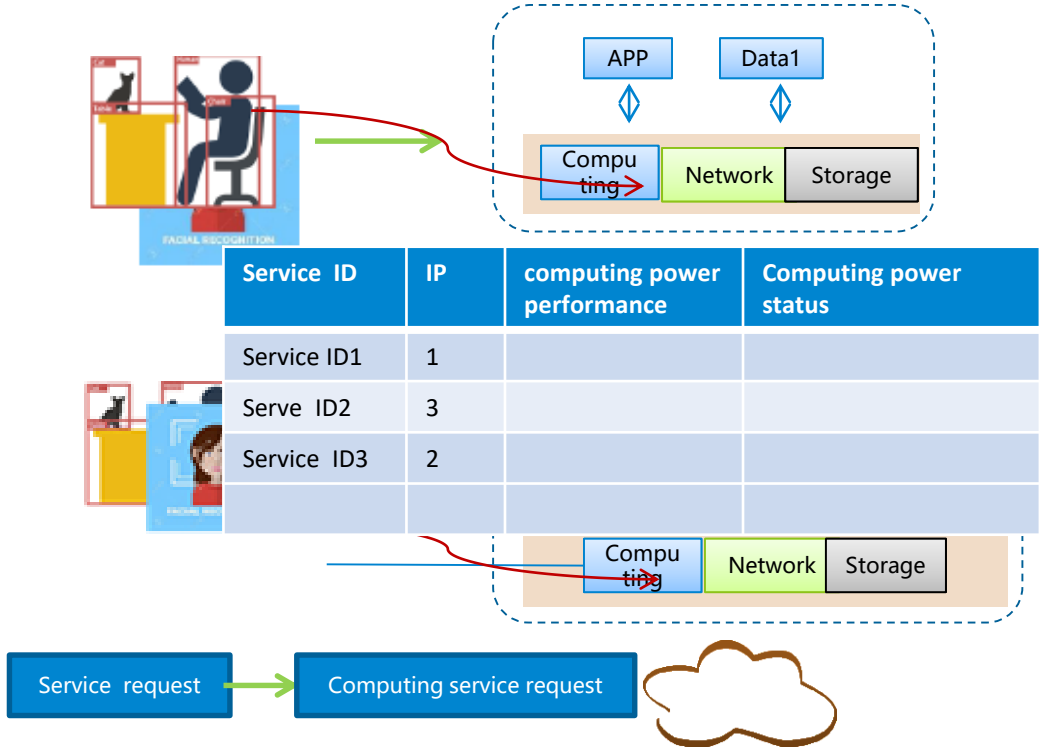
HCCR makes the computing resources aware by the network



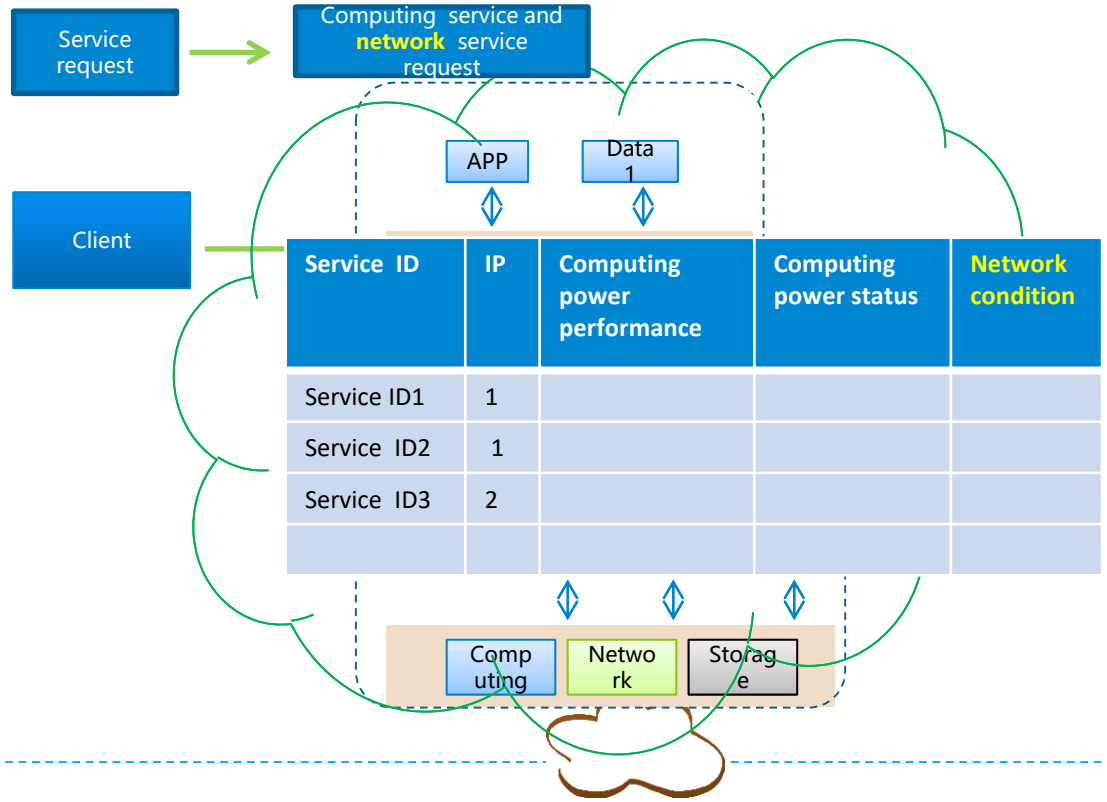
Hybrid Connectivity and Computing Routing (HCCR)

Inter-site(edge or cloud) computing power routing mechanism: 1.computing power scheduling; 2. computing and network resource coordination.

Computing power scheduling: 1. map the service request to the demand of computing resources; 2. schedule the task to the appropriate computing node according to the available power service resources.



Computing and network collaborative scheduling: according to service requests, dynamic task dispatch is performed in combination with network status and computing resource status.



Conclusion

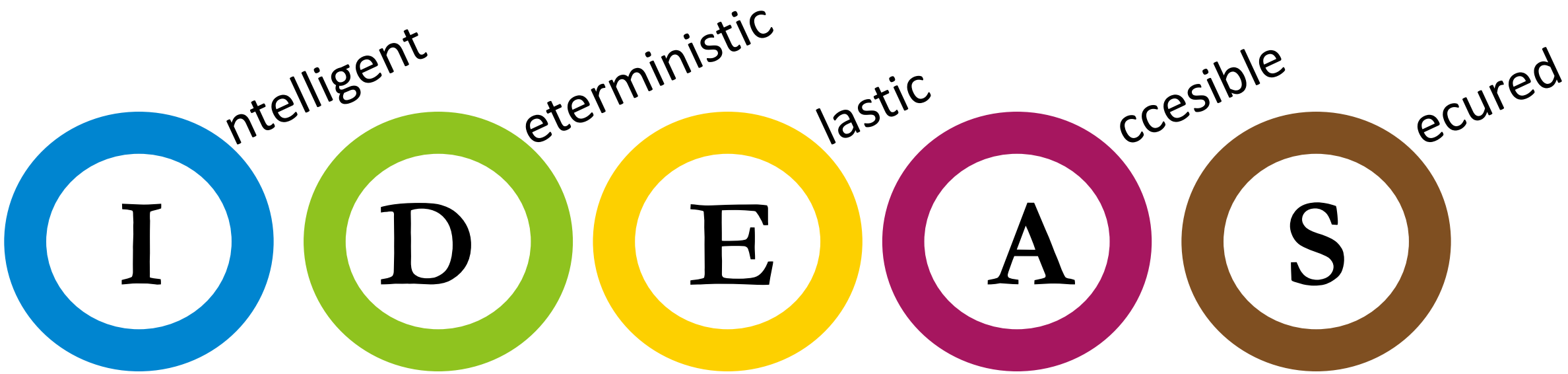
- New services such as 5G and industrial Internet pose new demands and challenges to IP network development.

- Three key demand:
  - Precisely Controllable Connectivity in Different Dimension
  - Widely Distributed Heterogeneous Computing
  - Edge/Cloud Coordination (Multi-domain)

- One proposed solution – High Precision Computing Network (HPCN)
  - High Precision Connectivity (HPC, Ongoing work such as TSN and Detnet in jitter dimension)
  - Heterogeneous Distributed Computing (HDC, FPGA, GPU, ASIC)
  - Hybrid Connectivity and Computing Routing (HCCR, Joint optimization)

Tell me about your IDEAS for the future

Design principle for future network defined in Architecture WG in Network 5.0 Alliance



*from "Whitepaper of Network 5.0 Industry and Technology Innovation Alliance"