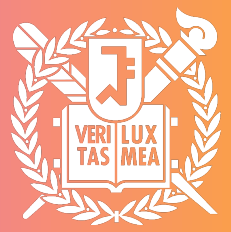


6G as Cellular Network 2.0

- A Networked Computing Perspective -

Mobile Korea 2023 – 6G Global 2023

2023/11/01

Kyunghan Lee

kyunghanlee@snu.ac.kr

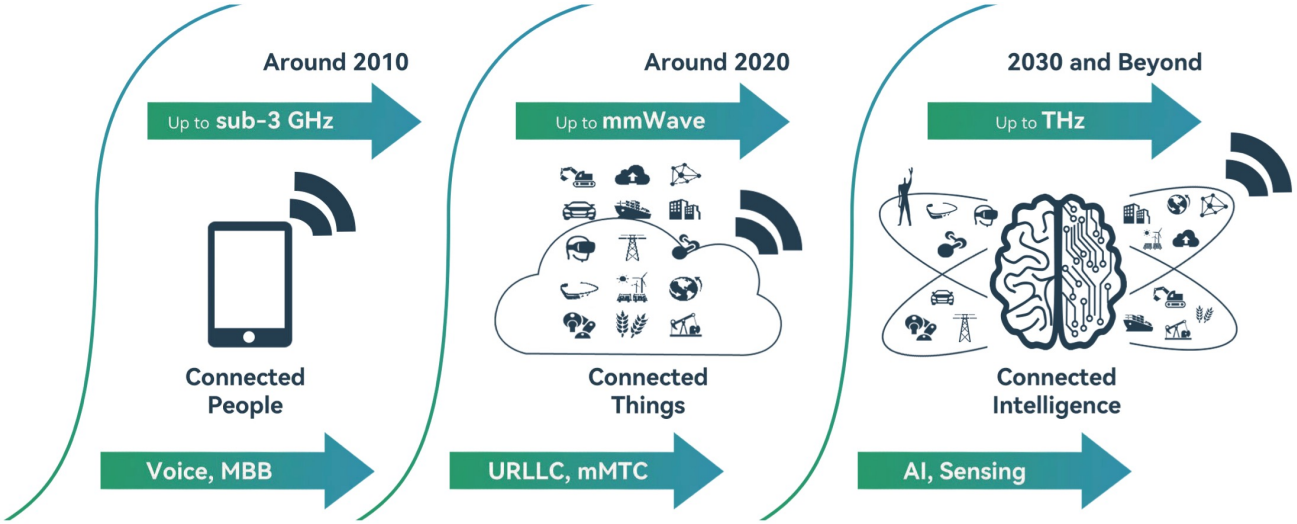
Networked Computing Lab.

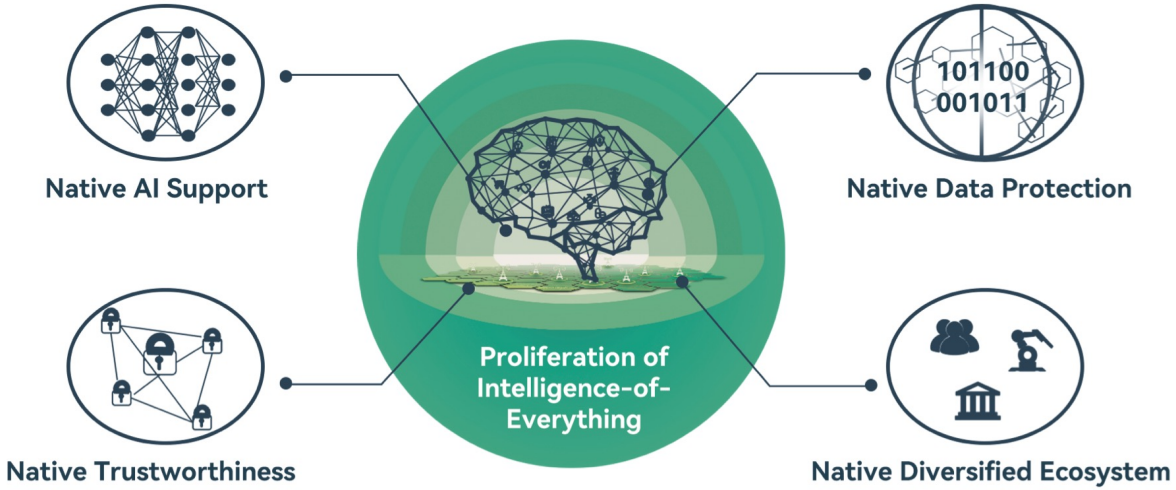
Electrical and Computer Engineering

Seoul National University

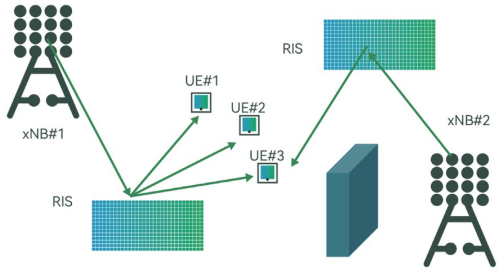
NXC LAB

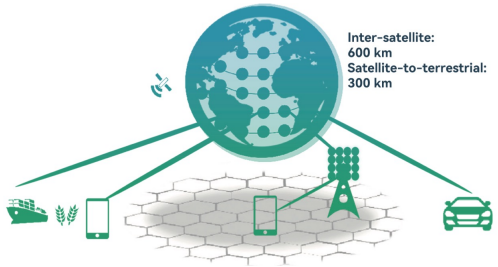
6G? An Ordinary Perspective

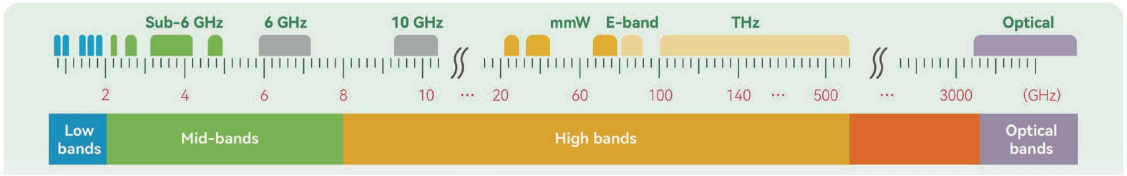




	3/4/5G	6G
Service	Connectivity Only	Connectivity and Sensing, AI as a Service
Networking	Public with Extended Private	Public Native and Private Native
Security	Encryption-based Security	Technology-based Trustworthiness
Algorithm	Analytic Only	Analytic + Data (AI)
O&M	Automated OA&M	Level 5 Native OA&M
Business	Networking Infrastructure	Networking & Computing Infrastructure
Coverage	Terrestrial Only	Integrated Terrestrial and Non-Terrestrial



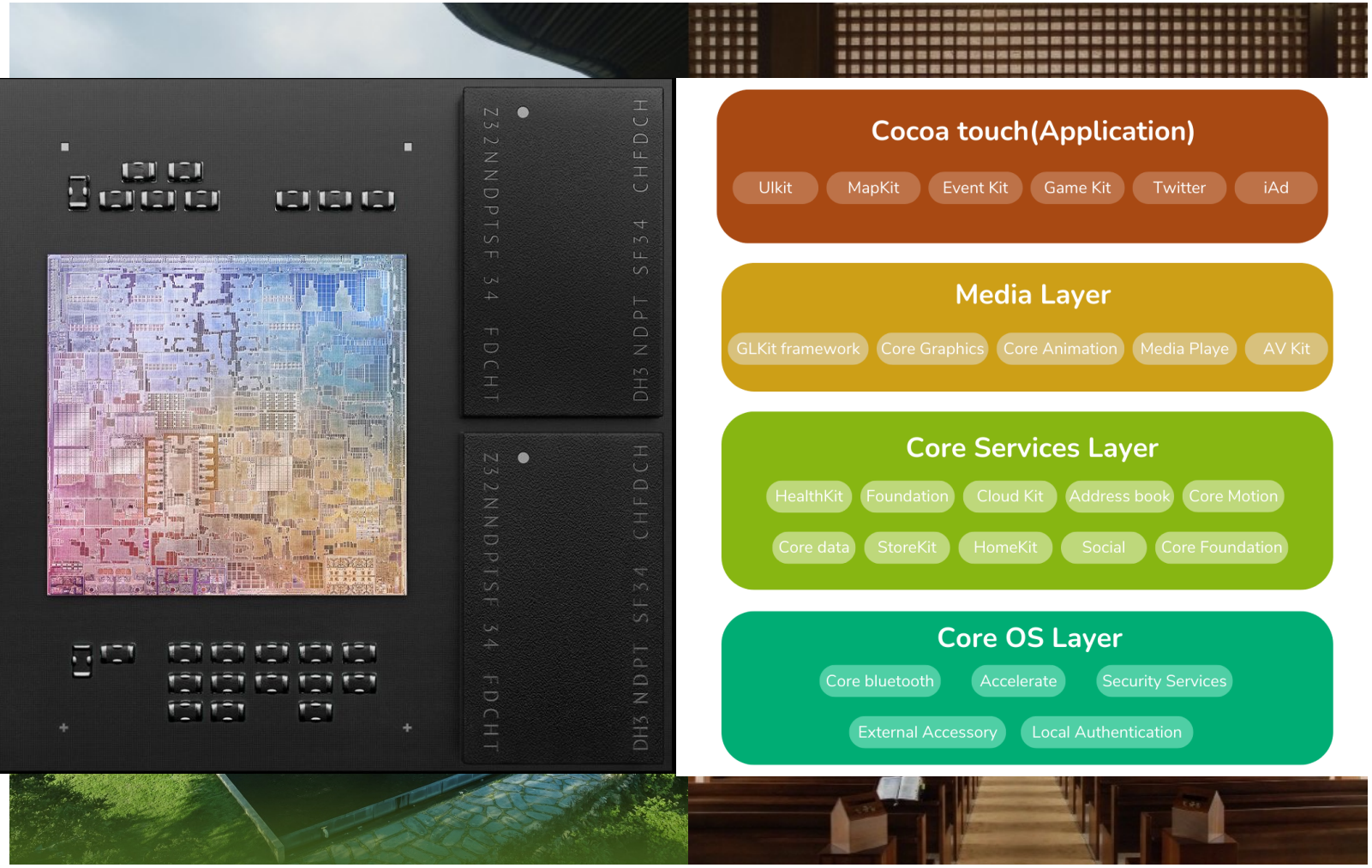


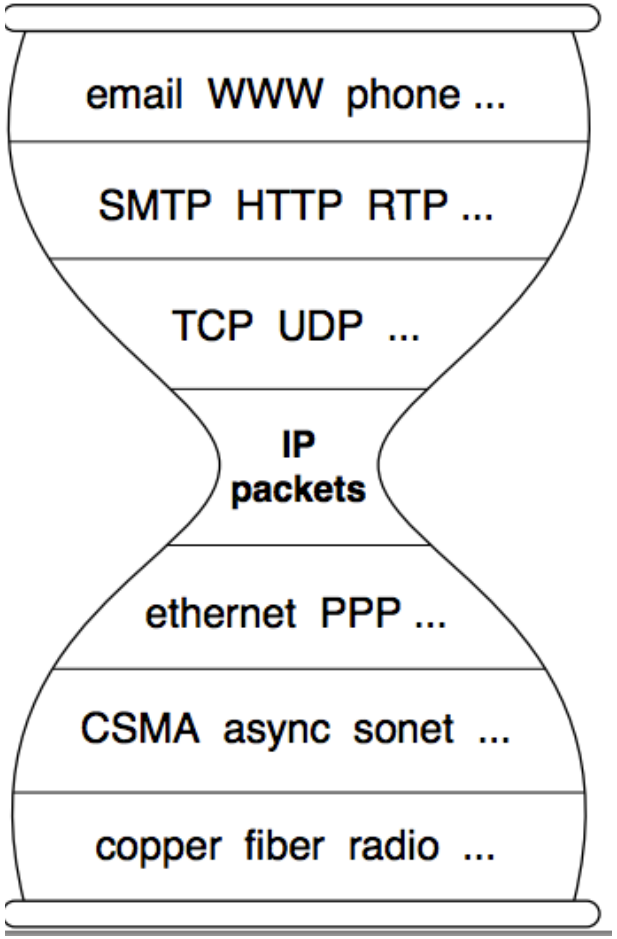


Huawei, "6G: The Next Horizon White Paper," white paper, 2022

Architecture

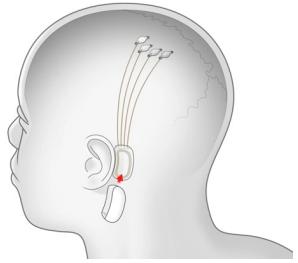
Architecture is the art and technique of designing and building, as distinguished from the skills associated with construction.
It is both the process and the product of sketching, conceiving, planning, designing, and constructing buildings or structures.

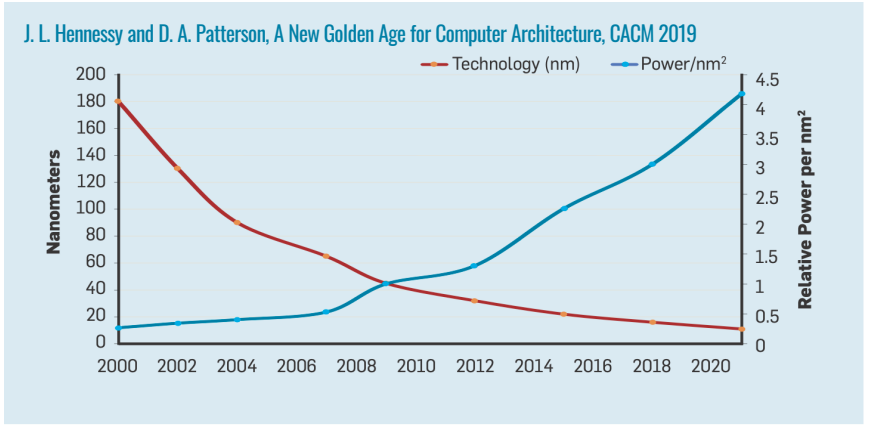




Computing over Networks (Connected Computing) is the Inevitable Future

Mobile/Wearable/Implantable Devices need ExoComputing Capability for Next-Generation Services





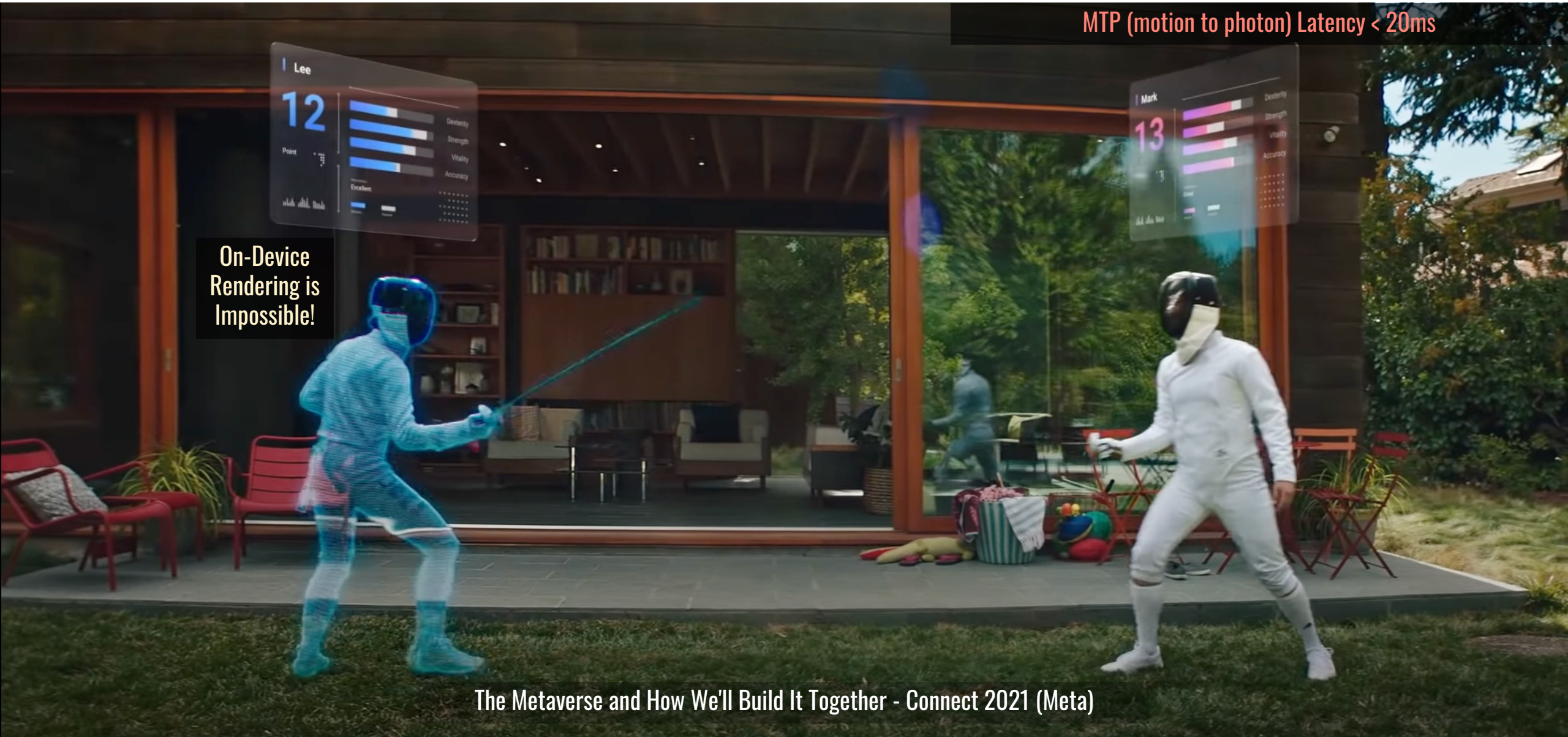
Performance-per-Watt improvement is slowing down.
(Dennard scaling came to an end in 2012)

Samsung Research, "6G The Next Hyper Connected Experience for All," white paper, 2020

Realtime Connected Computing is the Key to Human Augmentation

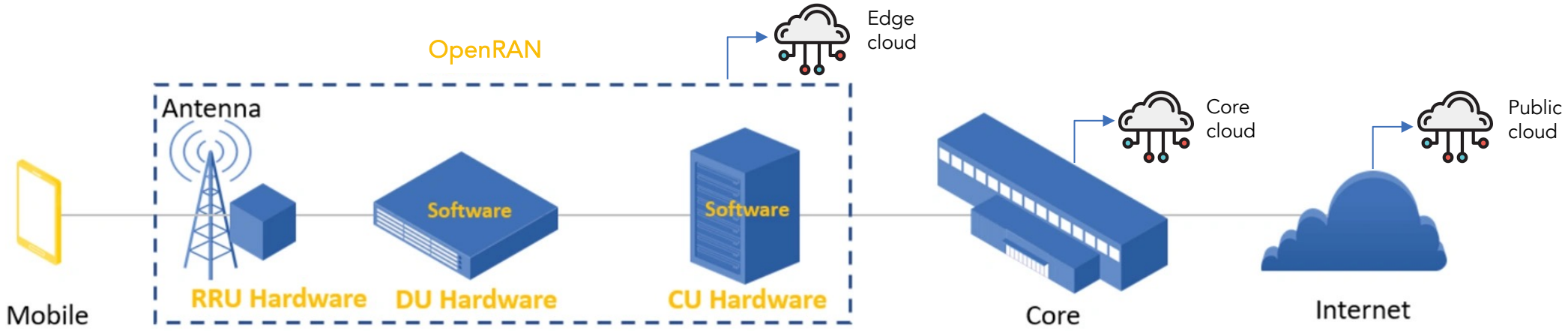
How Can we Enable Realtime ExoComputing (Connected Computing) over Wireless Networks?

Realistic Metaverse is an Example of Realtime ExoComputing



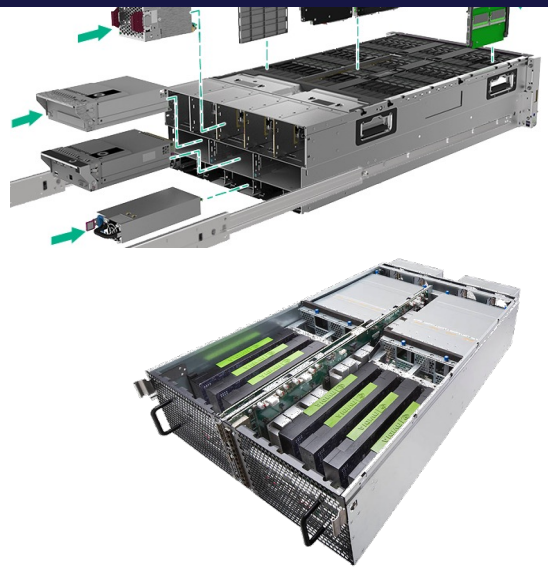
Offloaded AI Analytics is another Example of Realtime ExoComputing

Can Cloud Computing over 5G Networks Enable Realtime ExoComputing?



No. 5G Network is Too Unstable and Unpredictable.

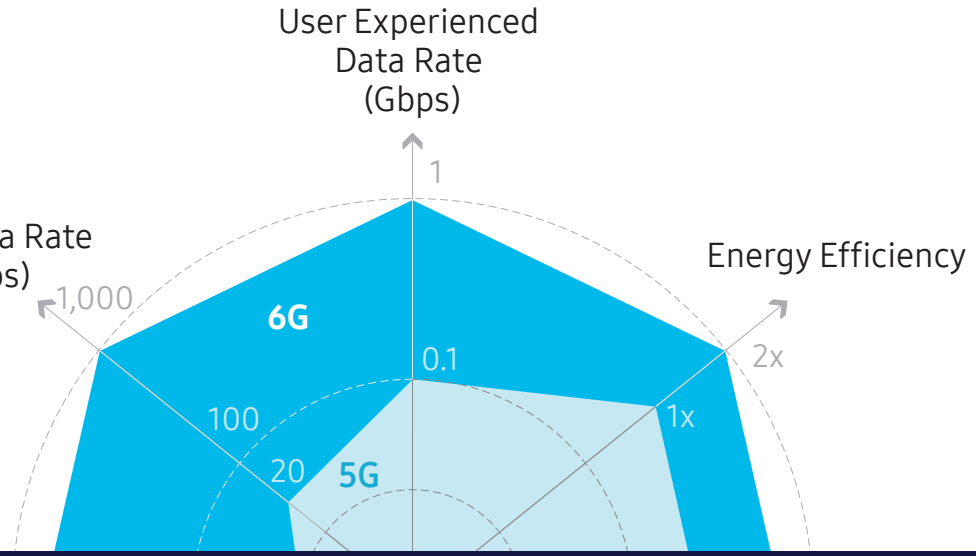
- Cooling Environments
- Data Center Facility
- Hardware Management
- Networking
- Open System Firmware
- Rack & Power
- Security
- Server
- Storage
- Telco
- Time Appliances Project (TAP)
- OCP Strategic Initiatives

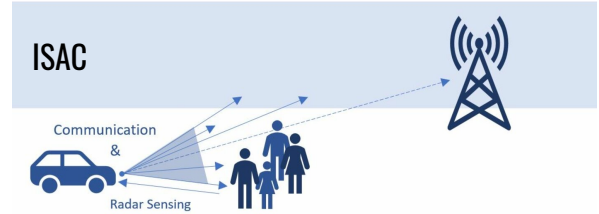


Would Deploying 6G Enable Realtime ExoComputing?

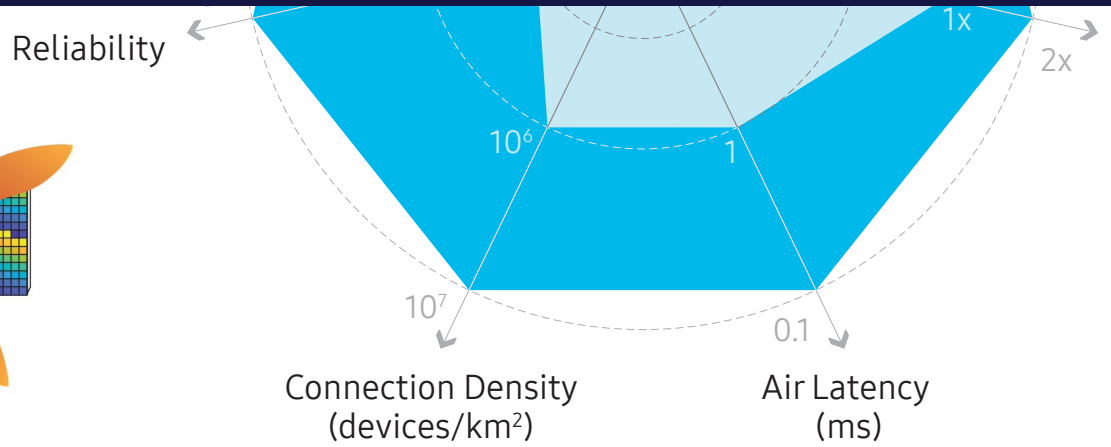
FSO

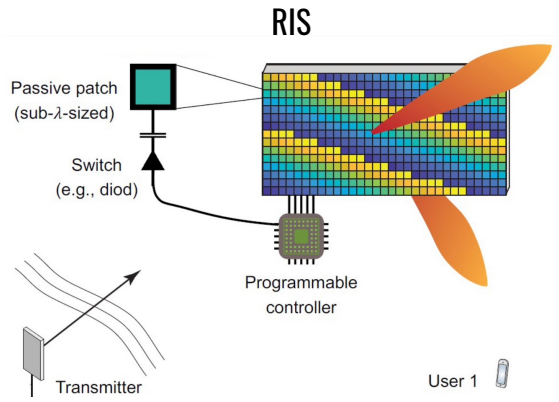
LEO Comm.

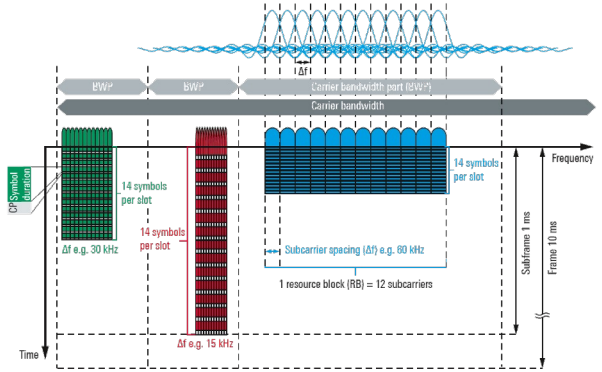




Unfortunately, Still No. Why?



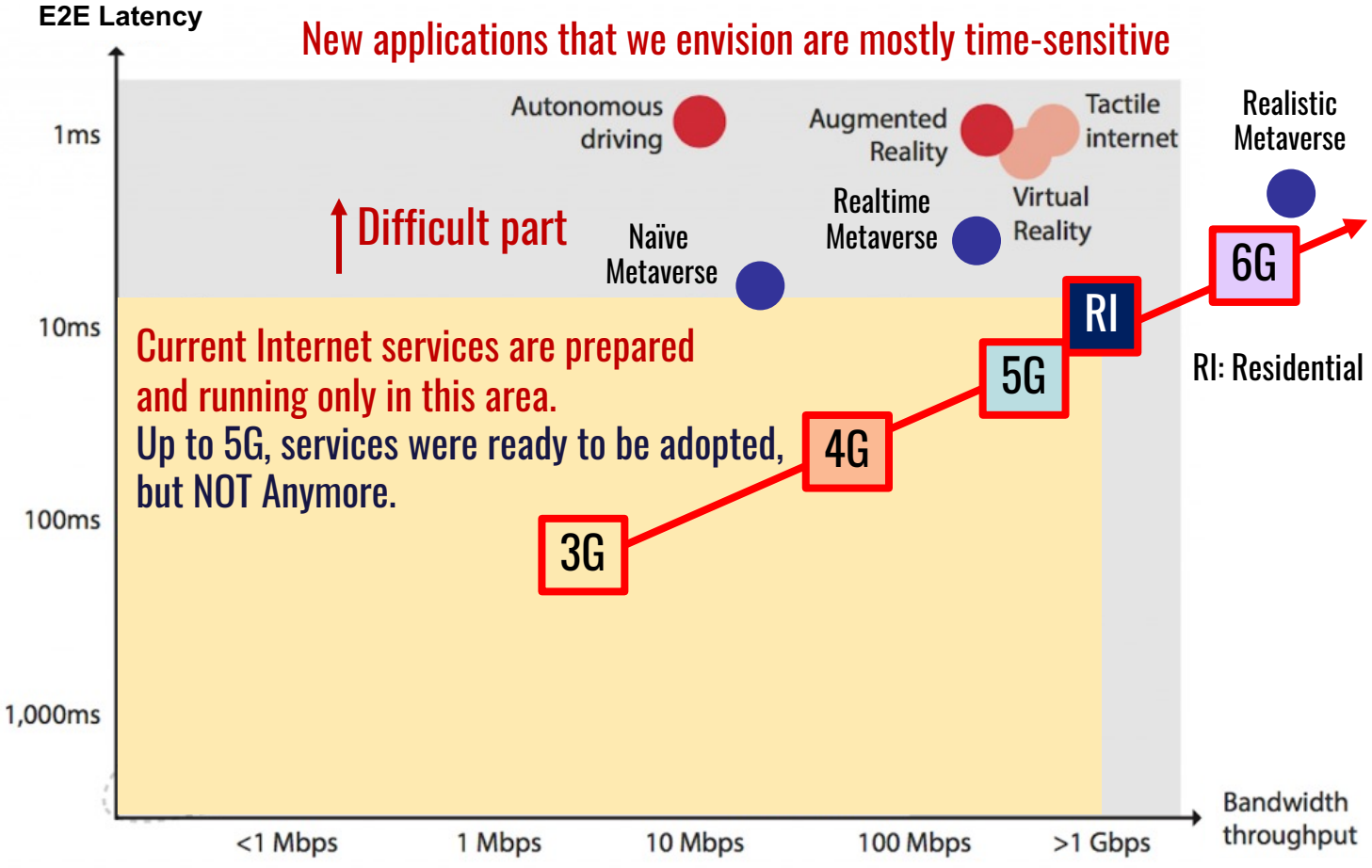




Numerology

Samsung Research, "6G The Next Hyper Connected Experience for All," white paper, 2020

What is Wrong with 6G Cellular Network System?



6G with MEC will surpass RI.

So, NEW Services will not be ready for 6G unless we prepare our own plan.

Mark Glimour, "5G – Latency: New use cases and the need for network slicing," InformaTech reports, Feb. 2017.

Low-level Evolution ≠ Perceivable Performance Improvement

Full potential of low-level evolutions can only be delivered to users via sophisticated software.
It's time to think about developing Cellular APIs for Realtime ExoComputing Services like Realistic Metaverse.

Realtime ExoComputing needs Cellular Network 2.0, not 6G

Cellular APIs can make the Network Application Performance Indistinguishable from the Performance of Applications Running on Device.

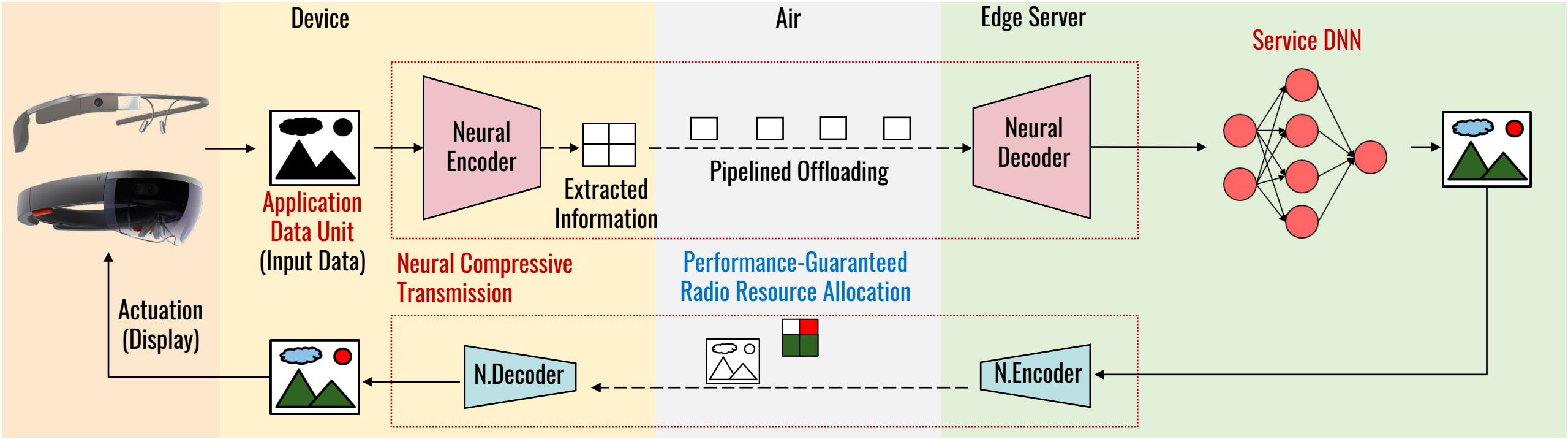
Cellular APIs ➔
Cellular Network 2.0
(Hyper-Connectivity)

From Connectivity to Hyper-Connectivity

Hyper-Connectivity:

New Connectivity that provides consistent application performance over the network even when the wireless channel therein fluctuates from device mobility.

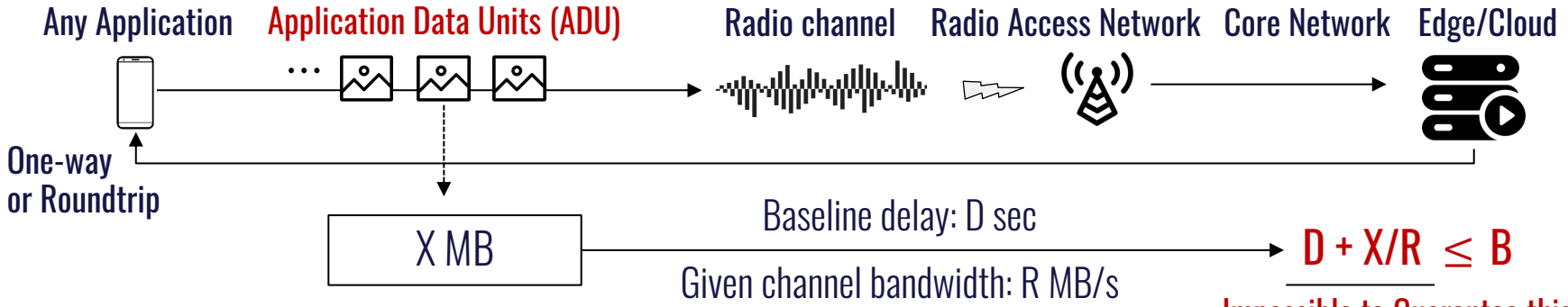
A Reference ExoComputing Service Model with Cellular Network 2.0



$$ADU.CT = T_{wait} = t_{en/decoding} + t_{tx\,up} + t_{computing} + t_{tx\,down} + t_{en/decoding} \leq T_{budget}$$

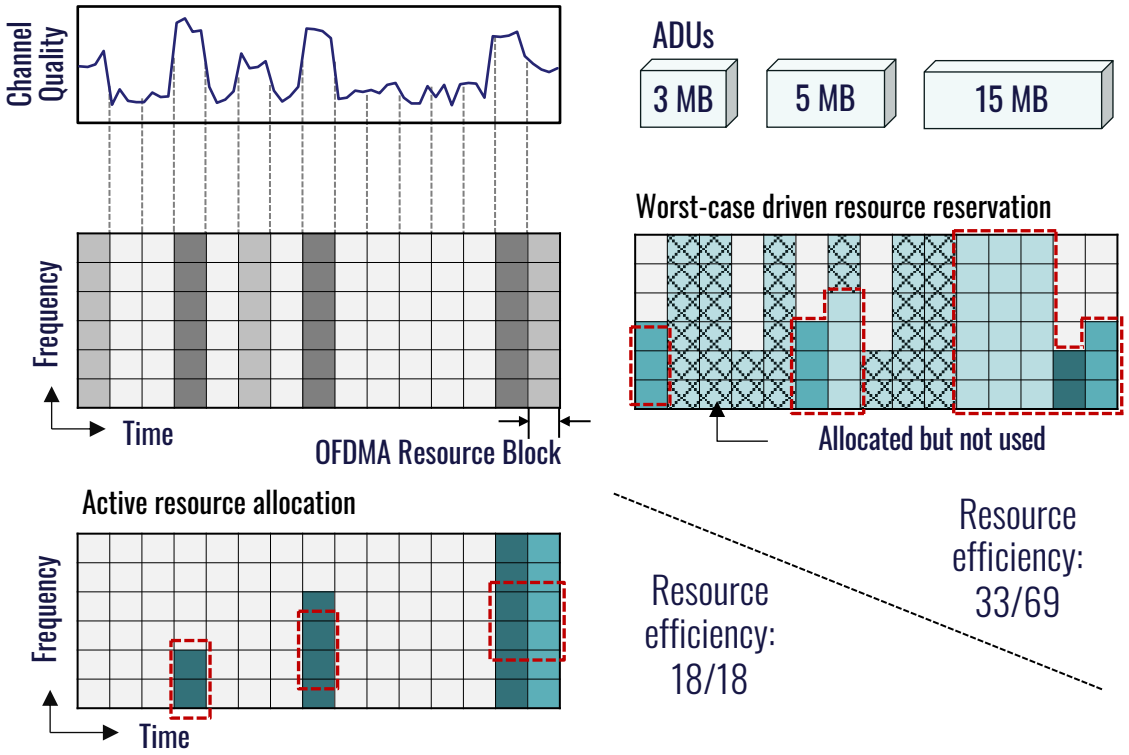
Networked Computing can Pipeline and Hide most of these Delay Components!

Enabling Cellular API (Hyper-Connectivity) ≈ ADU Completion Guarantee

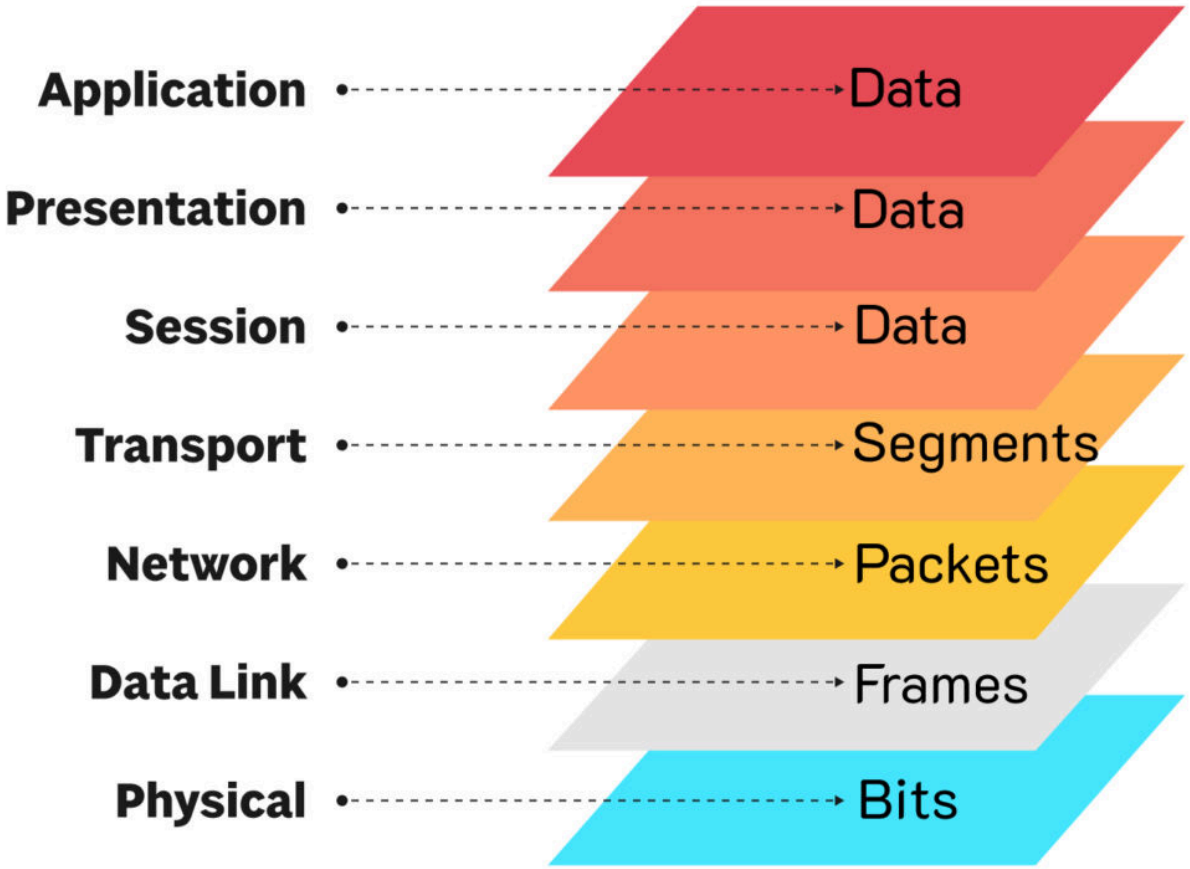


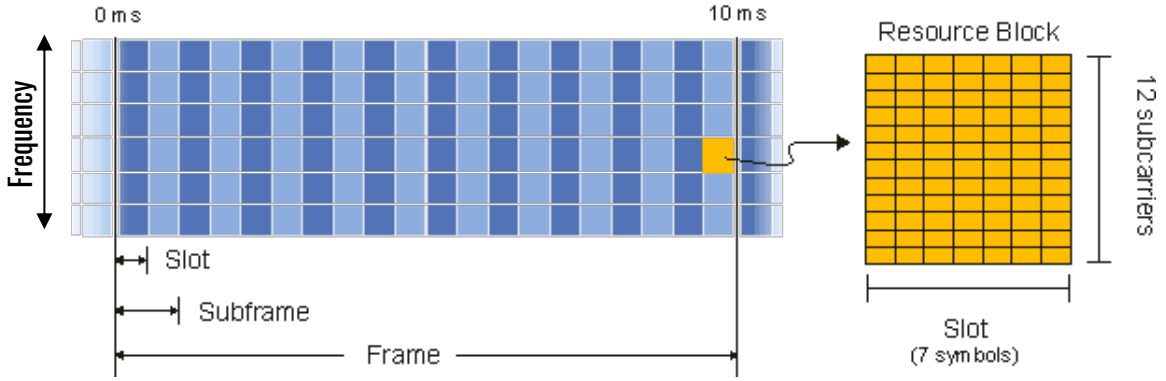
Conventional Approach
→ Network Slicing (Radio Resource Reservation)
- No guarantee
- Resource wate

Cellular API Approach
for Hyper-Connectivity
→ Active Networking
- Persistent guarantee
- New pricing needed (to service providers)

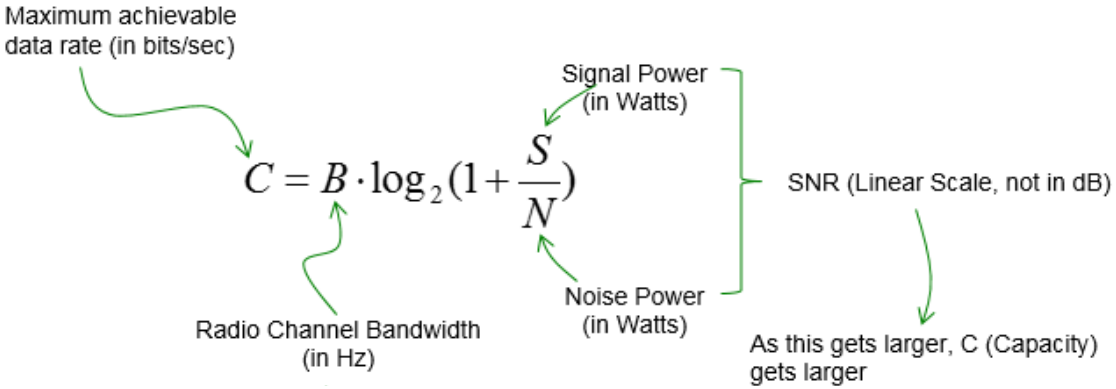


From the Conventional Bottom-Up Philosophy to a New Top-Down Philosophy



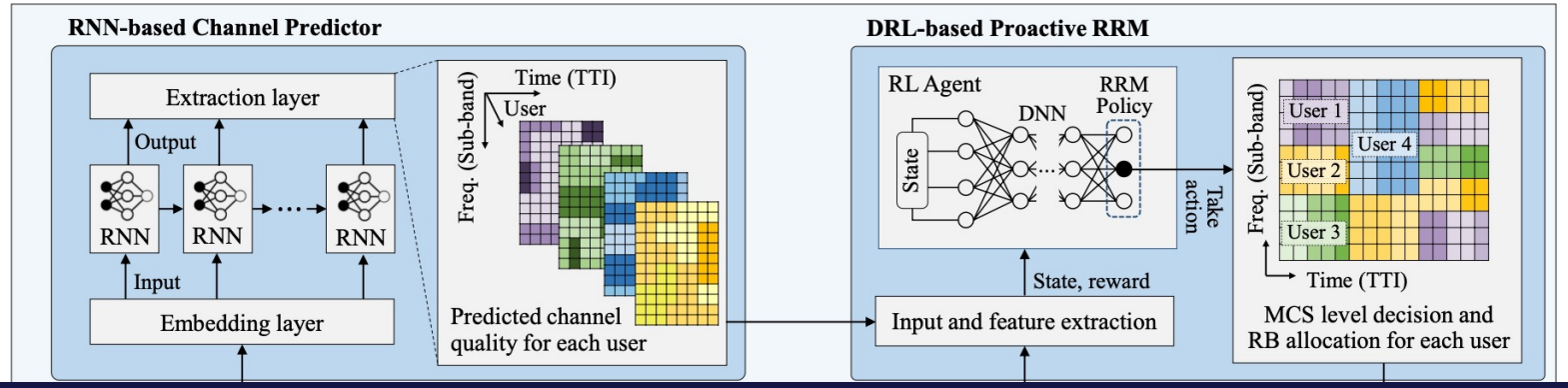


Shannon (Information) Capacity



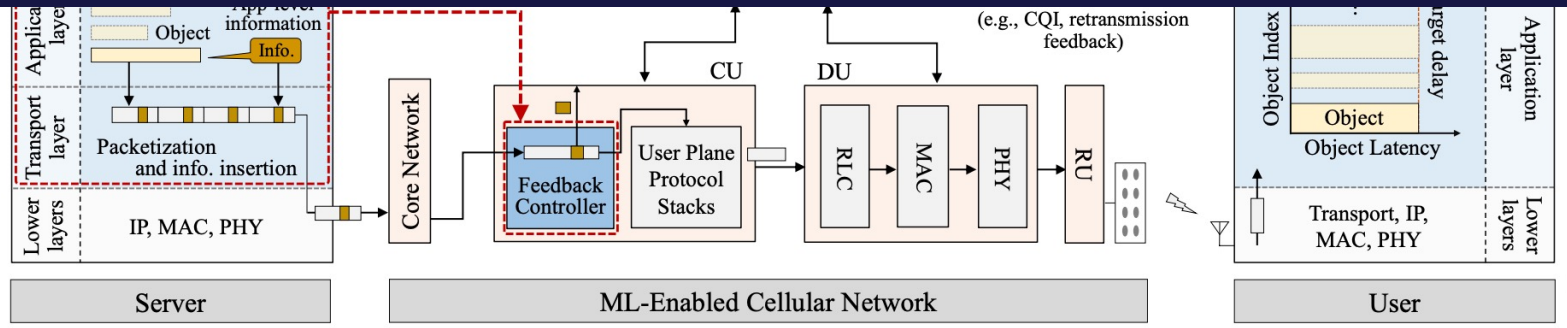
Cellular Network should be No More a Blackbox to Applications!

Towards Enabling Performance-Guaranteed Networking in Next-Generation Cellular Networks



Cellular Network should be able to Interact with Applications!

Cellular Network 2.0: Active Networking for Cellular APIs

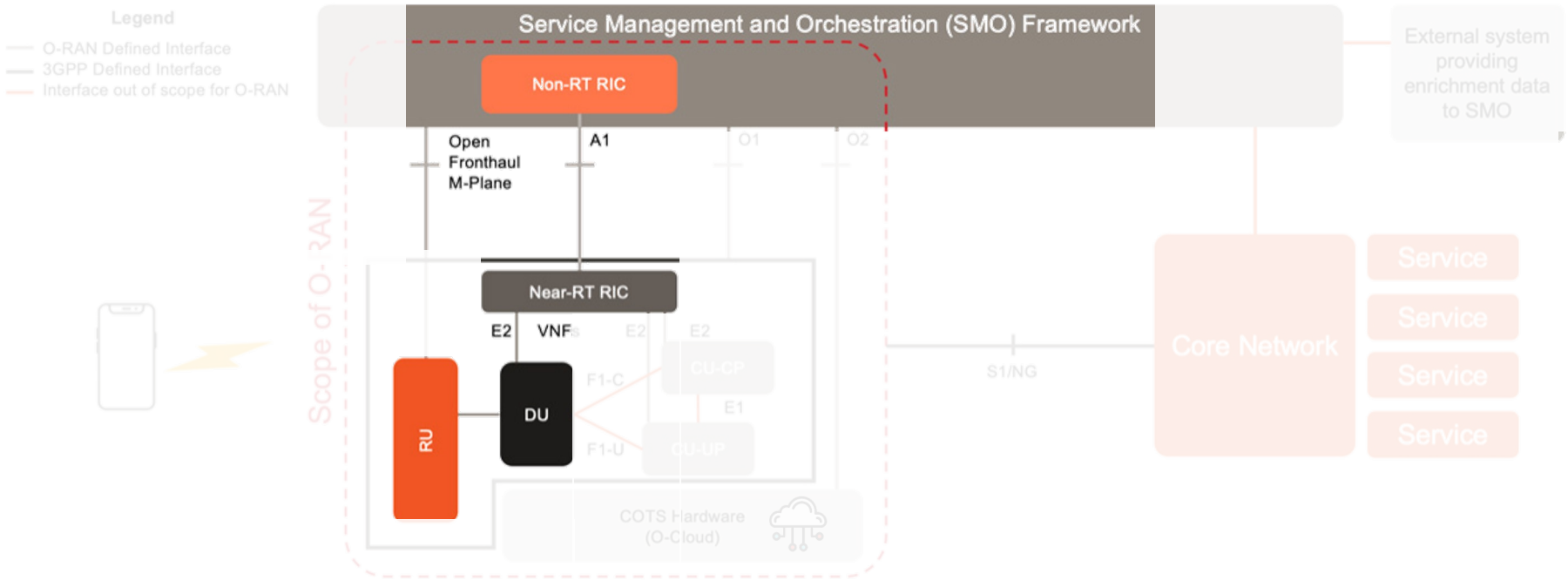


Junseon Kim, Byonghyo Shim, and Kyunghan Lee*, "Towards Enabling Performance-Guaranteed Networking in Next-Generation Cellular Networks," IEEE Communications Magazine, vol. 61, no. 1, pp.32-37, Jan. 2023

Active Networking Can Make Cellular Network a Whitebox to Applications!

Cellular API supported by
Softwarized Network Orchestrator

- RAN time budget
- Core-network time budget
- Computing time budget



True Benefits of Cellular Network 2.0 for Realtime ExoComputing

Hyper Connectivity to Applications from Cellular APIs

Infinite Computing Capability to Network Devices

Services running over CN2.0 becomes indistinguishable from services running on device.

Guaranteed ADU CT

Network application programming becomes straight-forward.
(socket programming was a disaster.)

```
frame = capture( sensor );
display( Cellular_API( frame, destination, service deadline ) );
```

Pricing/incentive model for Cellular_API()
* Difficulty-based

1) $CT_{UL} \leq T_{budget}$

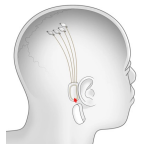
2) $CT_{DL} \leq T_{budget}$

3) $CT_{RoundTrip} \leq T_{budget}$

Is Hyper-Connectivity a Matter of Choice?


Type 2


Type 1


Type 0

Computing over Hyper-Connectivity

vs.

On-device Computing

Near Real-time
AI/ML Analytics/Computing
Services (e.g., GPT-4)

No. Hyper-Connectivity is Indispensable! However, Is Networking Really Power-Efficient?

Higher Performance ← Higher Power


RTX3090
(350W)

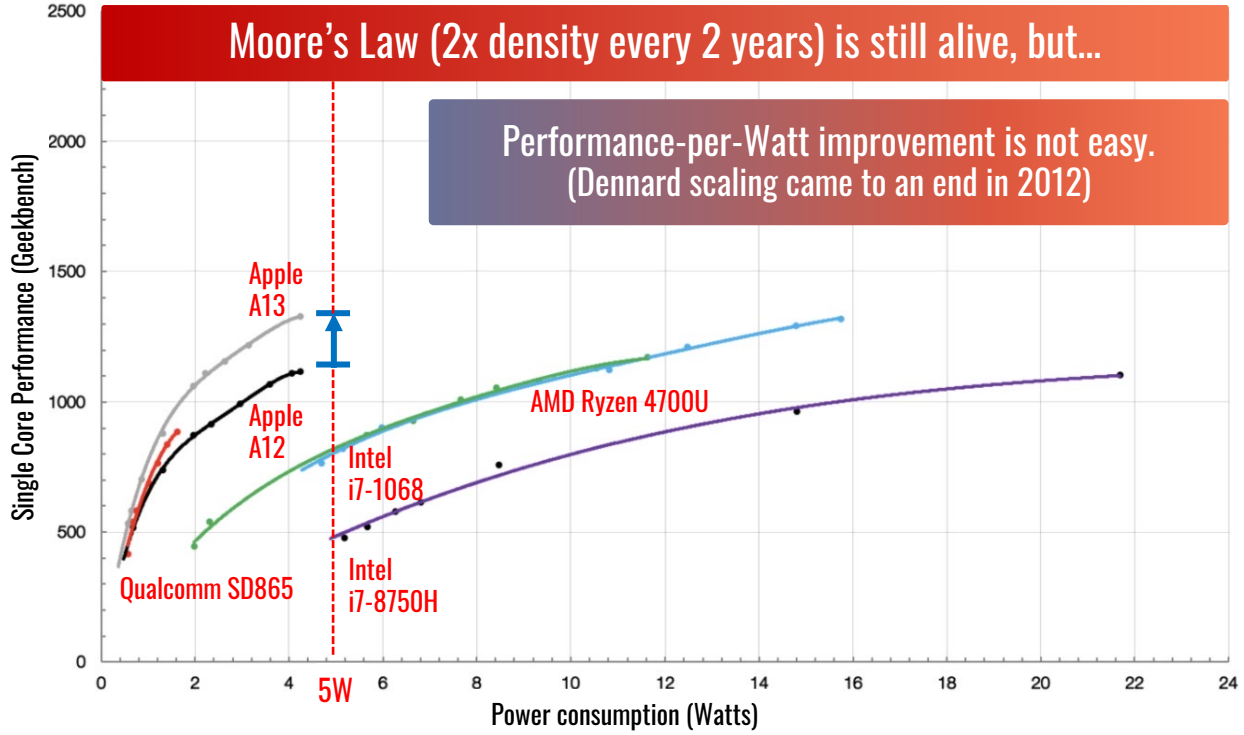

RTX2080
(200W)

Thermal & Battery Capacity Limited (Type 1 & 0)


SD888
(5W)

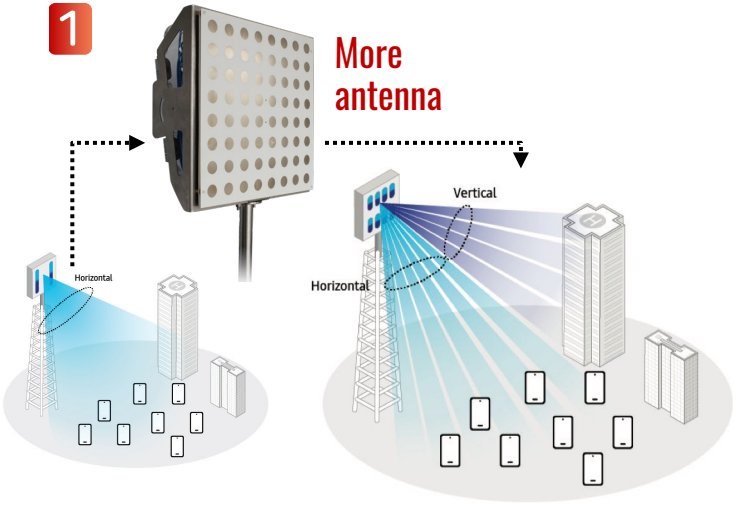
Moore's Law (2x density every 2 years) is still alive, but...

Performance-per-Watt improvement is not easy.
(Dennard scaling came to an end in 2012)

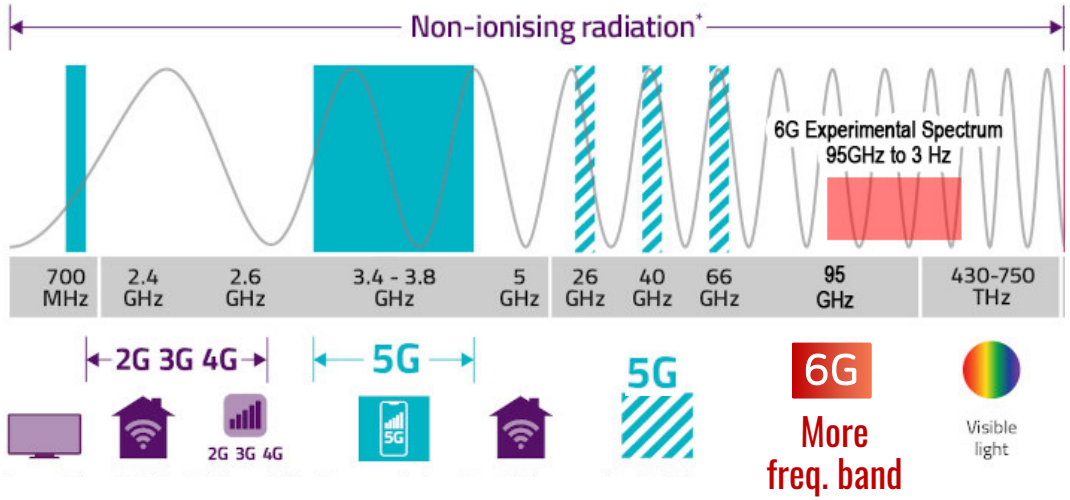


Is Hyper-Connectivity Achievable with Limited Power Consumption?

1

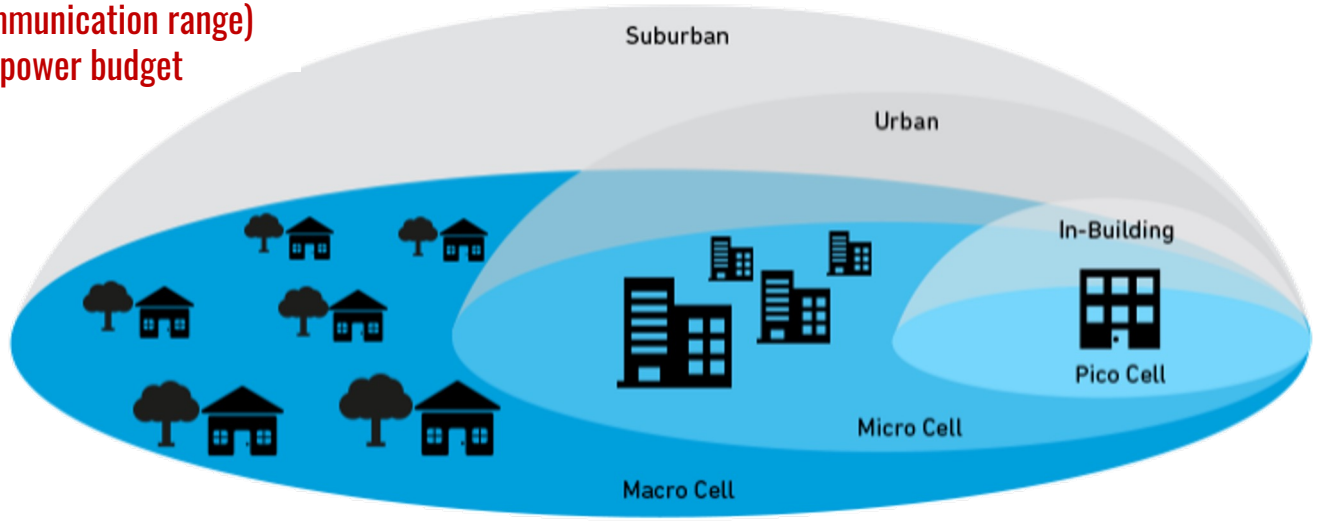


2

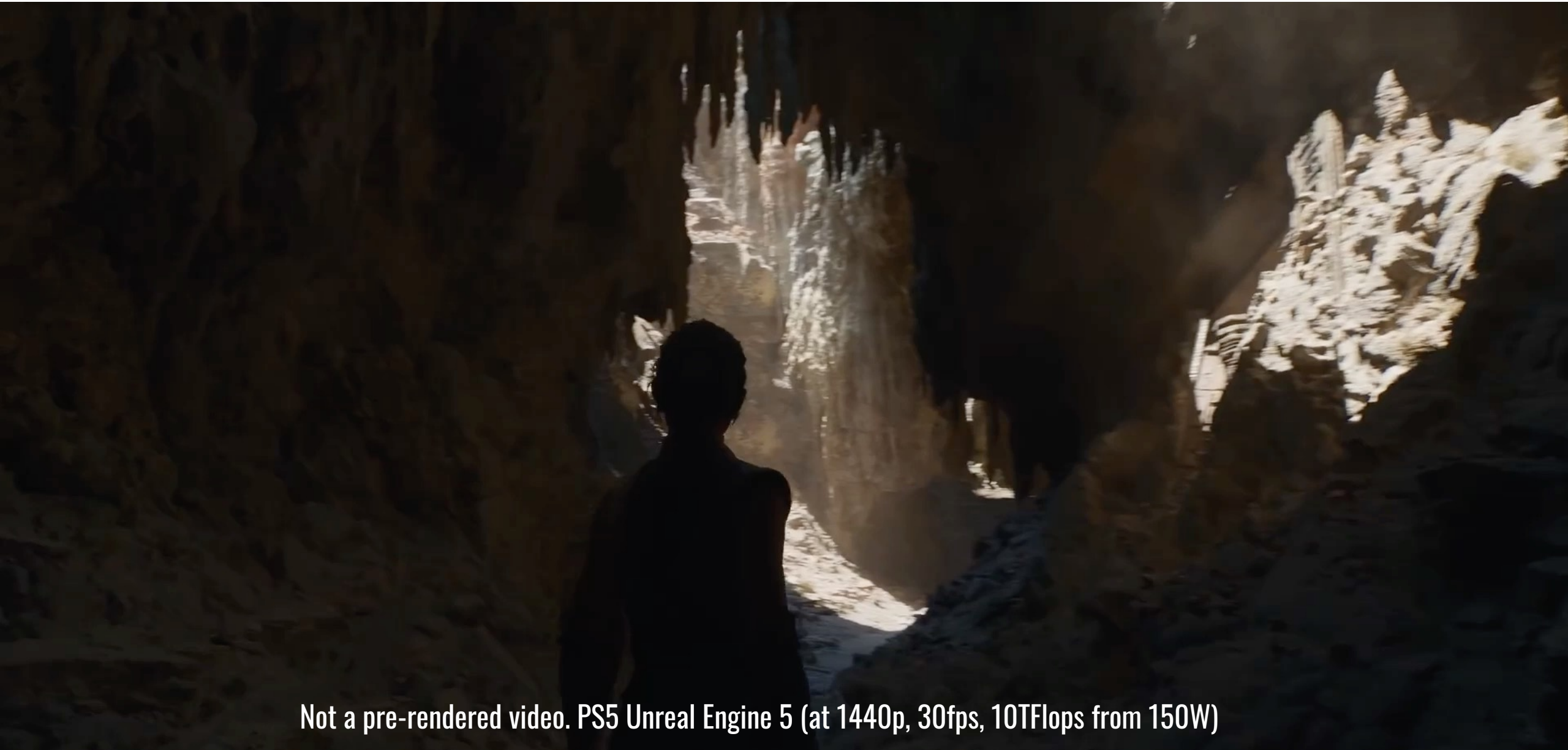


Unlike On-device Computing, We have Multiple Options to Limit the Communication Power!

3 Ultra dense deployment
(shorter communication range)
to meet the power budget

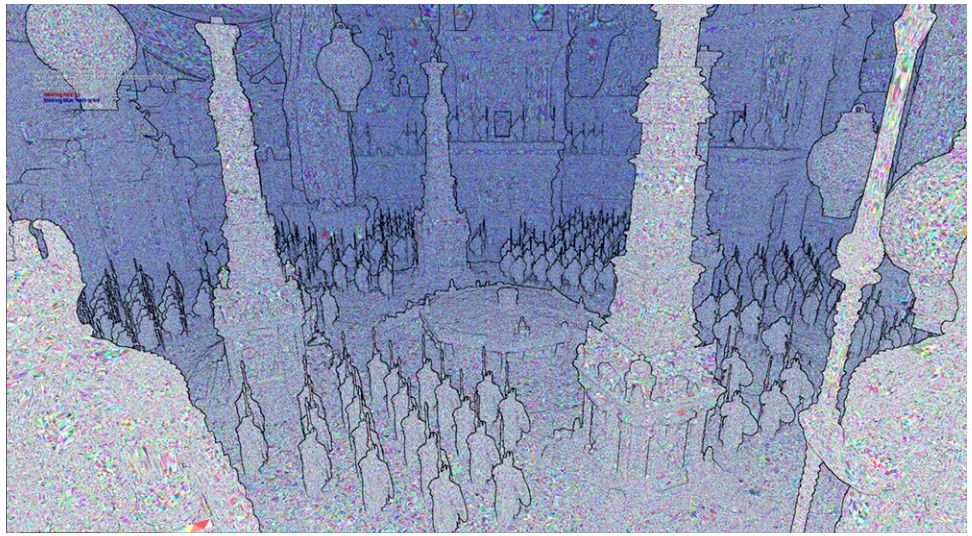


ExoComputing with Hyper-Connectivity vs. On-device Computing



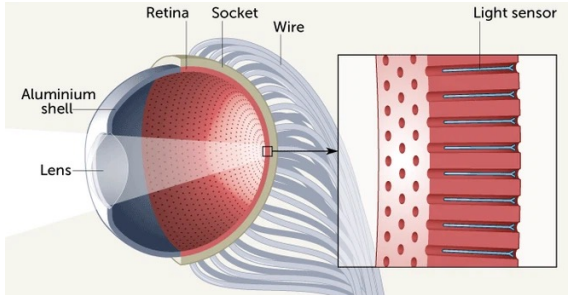
Not a pre-rendered video. PS5 Unreal Engine 5 (at 1440p, 30fps, 10TFlops from 150W)

ExoComputing with Hyper-Connectivity vs. On-device Computing



1440p Rendering on Device (3.5 MP)

30 fps - 33 ms (budget that a frame is rendered)
150 W (10 TFlops, about 16 billion triangles)



Human Eyes Spec
~ 2 x 63 MP

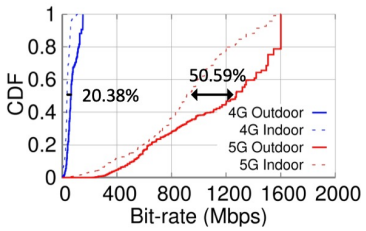
RD: x(6x6x6)
SR: x(6x6)

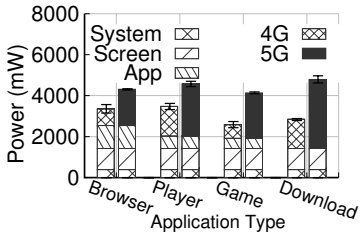
32 KW

1440p Streaming from Edge (3.5 MP)

14 MB (BMP, 2560x1440x32 bits)
<1.4 MB (JPEG, BPG, Web-P)
9.3 ms (1.4 MB / 1.2 Gbps)
2.5 W (1.2 Gbps)

5 W @ 25ms
(50 MB / 16 Gbps)



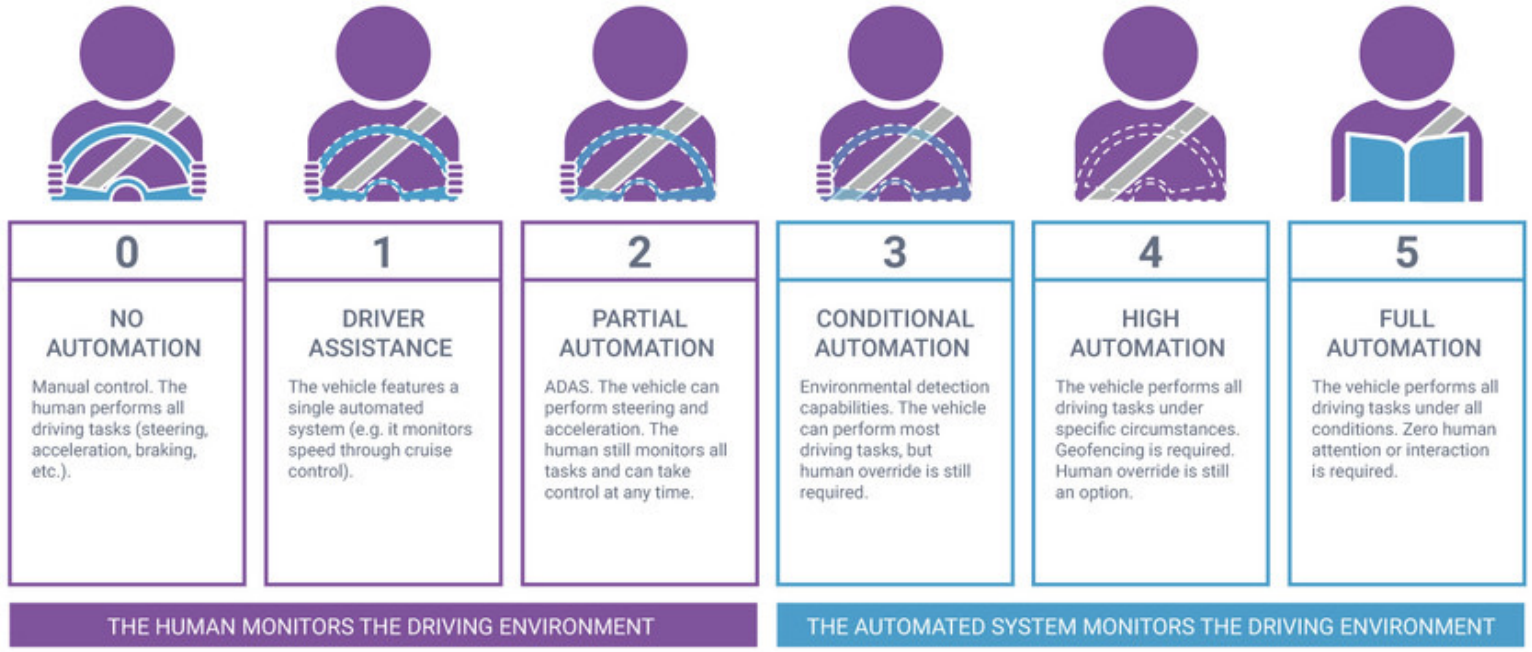


Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption, ACM SIGCOMM 2020

ExoComputing with Hyper-Connectivity vs. On-device Computing

A Roadmap of Enabling Hyper-Connectivity (6G & Beyond)

Levels of Driving Automation



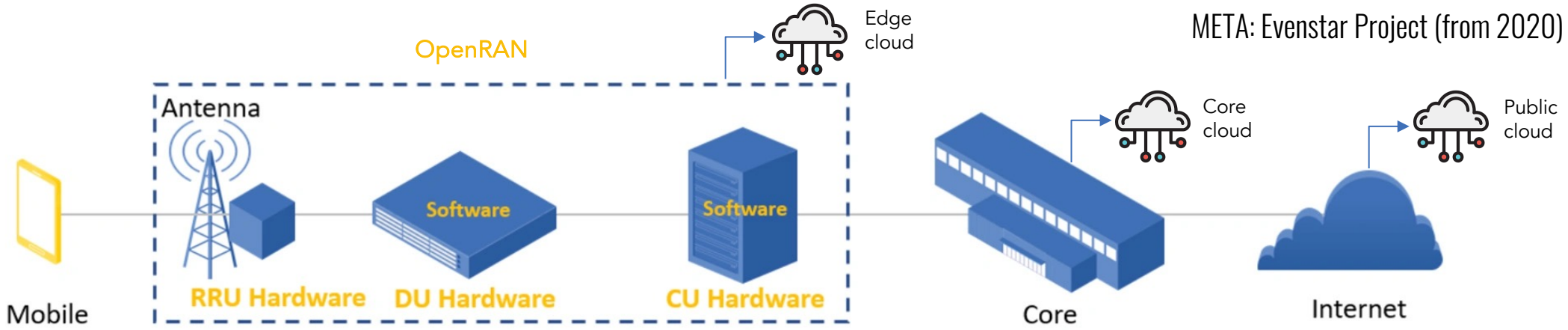
Levels of Performance Guarantee

(per-Region Certification)

0	1	2	3	4	5
Best effort (Zero Guarantee)	Performance Assistance Bandwidth guarantee	Partial Performance Guarantee ADU CT guarantee at static env.	Conditional Performance Guarantee ADU CT guarantee at low mobility	High Performance Guarantee ADU CT guarantee at high mobility	Full Performance Guarantee ADU CT guarantee at 3D mobility

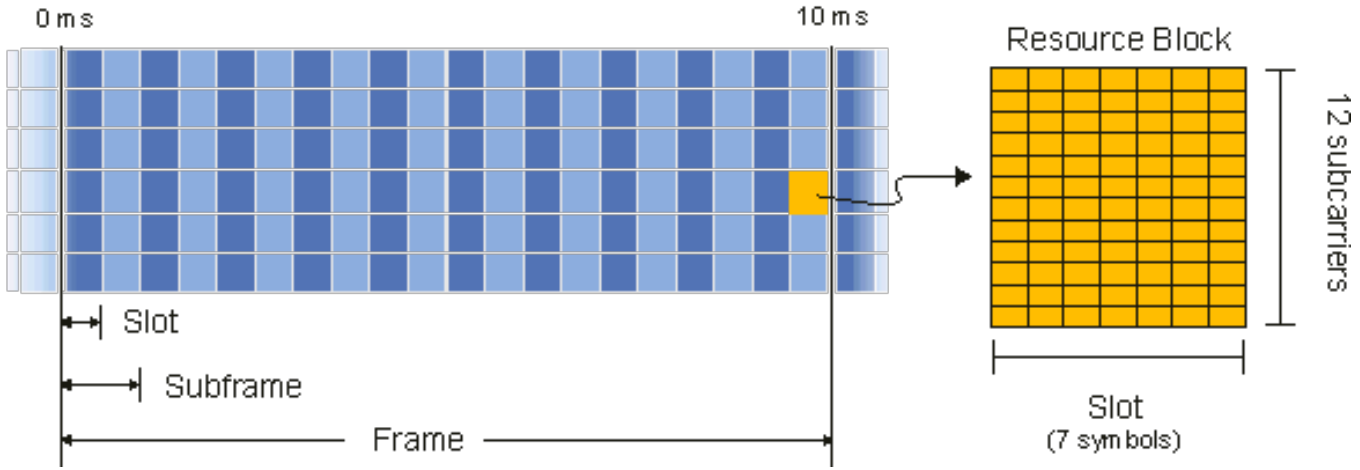
5G 6G 7G

Emergence of a New Networked Computing Service Operator



$$JobCT = T_{wait} = t_{en/decoding} + t_{tx\ up} + t_{computing} + t_{tx\ down} + t_{en/decoding} \leq T_{budget}$$

New Business Question: What is the Value of One Radio Resource Block in Terms of GPU Cores?



≈

Thank You

kyunghanlee@snu.ac.kr
nxc.snu.ac.kr