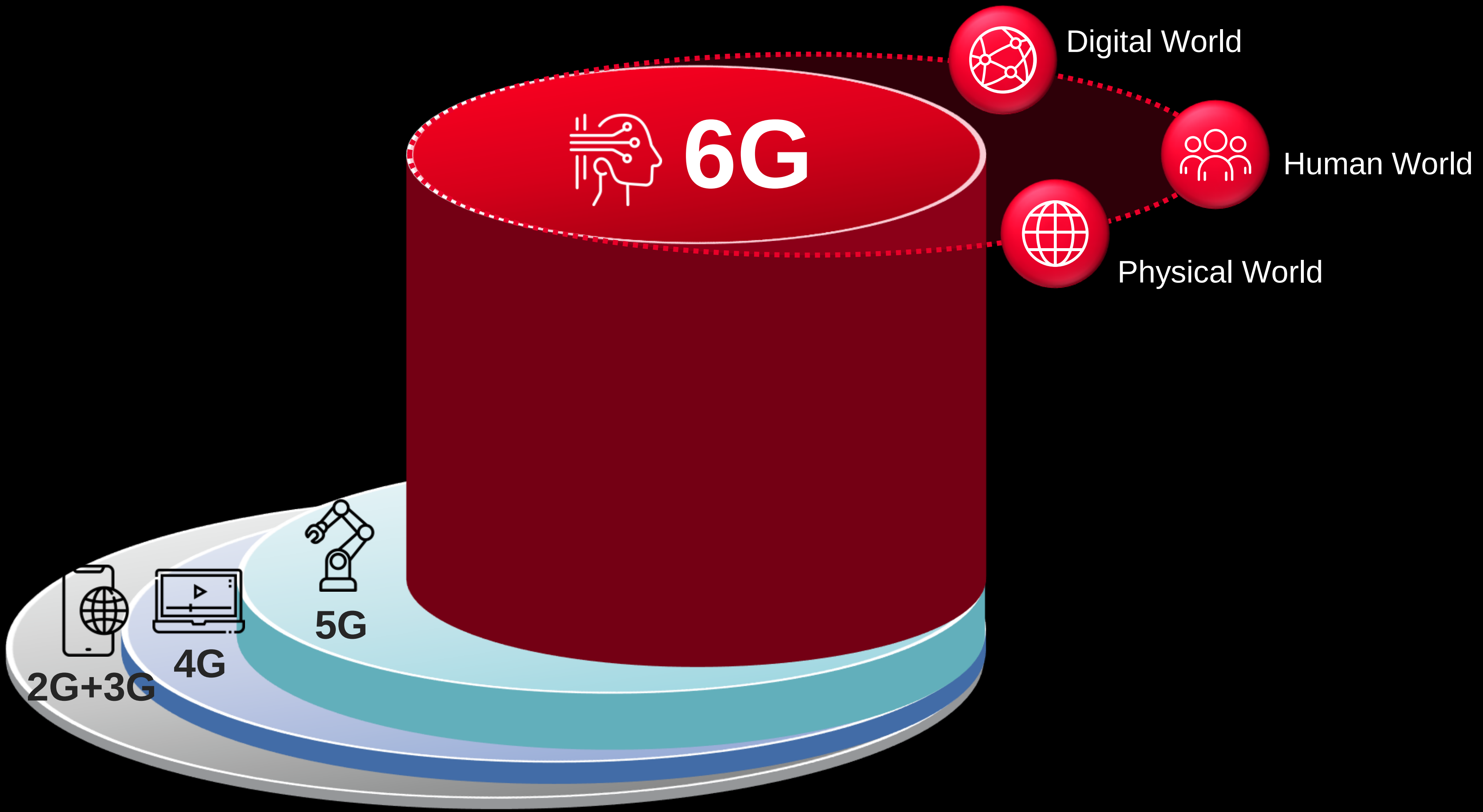
6G Readiness: Technology Insights for Tomorrow

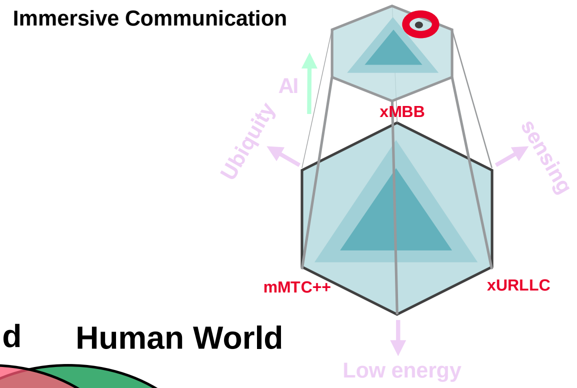
Mombasawala Mohmedsaeed
May 14, 2025
Bharat 6G 2025
New Delhi

6G Will Connect the Physical, Digital, and Human Worlds

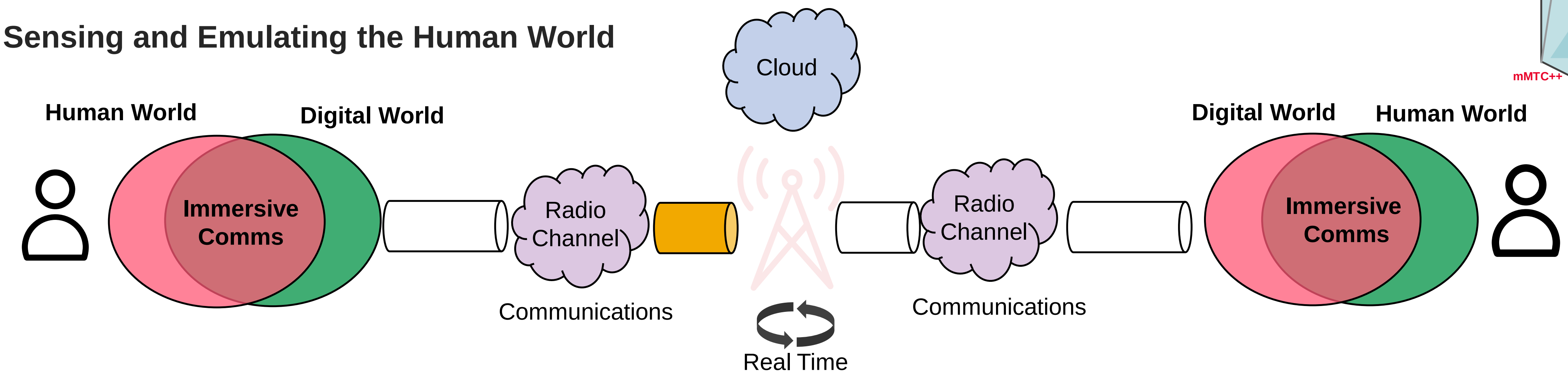


6G Monetizable Use Cases

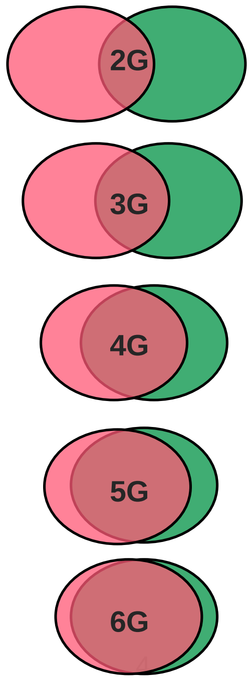
Immersive Communication – 2G to 6G Evolution



Sensing and Emulating the Human World



"G"	Sensor	Frequency Req	Tput Req	Latency Req	Comments / Technology Requirements
2G	Audio	900MHz	8-32kbps	<100ms	
3G	Video	900MHz / 1.8GHz	2Mbps	<100ms	
4G	Location	900MHz / 1.8GHz	100's bps	<1s	
5G	XR	3.5GHz	100Mbps+	<20ms	
	High accuracy positioning	3.5GHz	1kbps	<1ms	Need wide bandwidth signal to cross correlate for delays and location
6G	Immersive cloud XR	3.5GHz	10Gbps+	<10ms	High resolution immersive XR Watch live football game from the referee's viewpoint FR3 development
	Haptic information	n/a		0.1ms	Touch, motion, vibration Low latency communication
	Integrated Sensing and Communication	Mid band? THz?	?	?	Mid band / THz Technology development
	Holographic display	THz?	>1Tbps	<1ms	



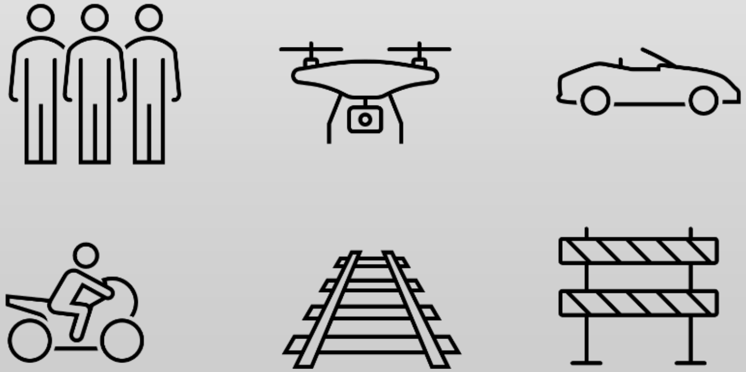
6G- New Market Segments

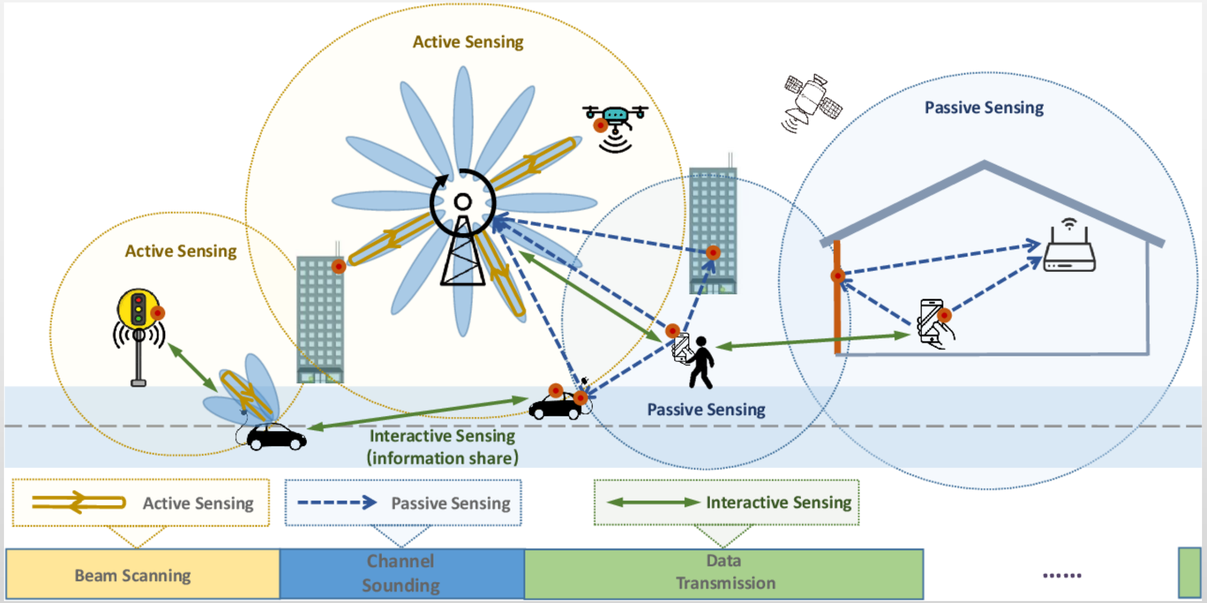
Integrated sensing and communications (ISAC)

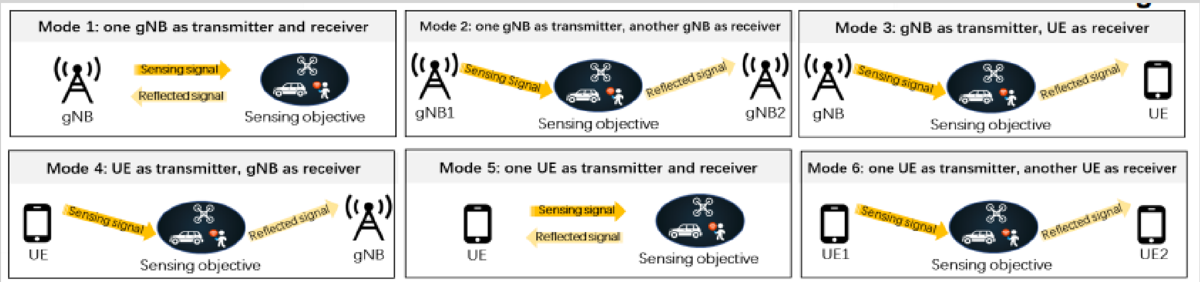
https://api.semanticscholar.org/CorpusID:234095582

ISAC

R19 SI Focus: Define channel modelling aspects to support object detection & tracking







KEYSIGHT

Four Key Technology Areas Driving 6G

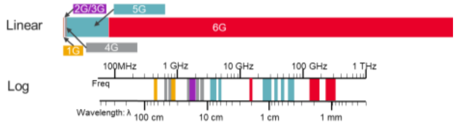
New Spectrum Technologies

Artificial Intelligence and Machine Learning

Digital Twins

New Network Topologies

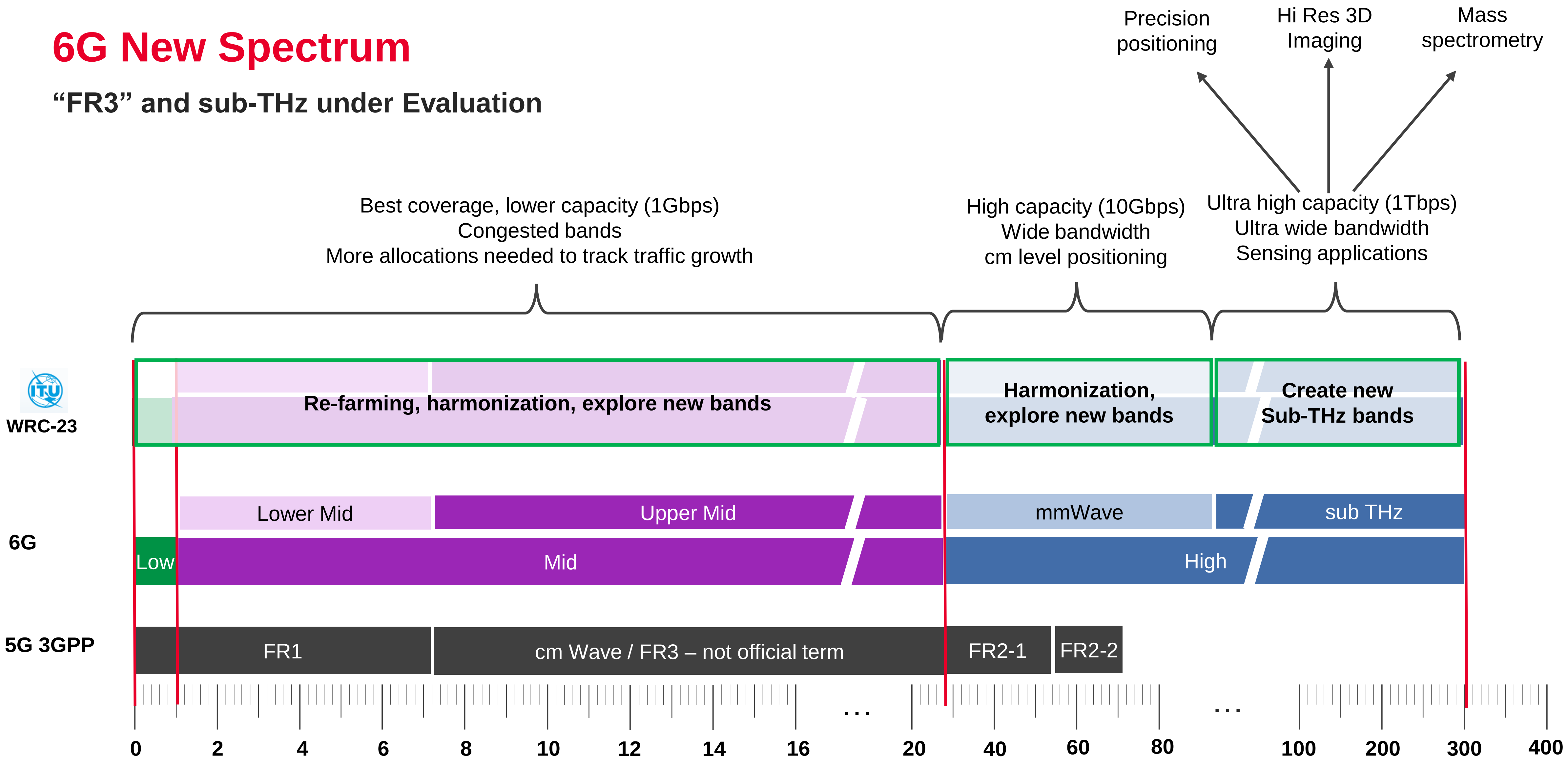
6G Candidate Spectrum: Specifics



	6G Research Topics										Mobile Regulatory Situation	Technical Challenges
	NOMA	Waveforms	Channel Coding	Unlicensed/WiFi	Advacned MIMO	Satellite	Mobility/Coverage	Radar/ISAC	PA & LNA	Antennae		
<7GHz	✓	✓	✓	✓	✓	✓		✓	✓	✓	• Moderate changes ongoing e.g. 3.4 GHz and 6-7 Ghz. Most allocations/auctions complete.	• Coverage • Spectral Efficiency
7-16GHz		✓	✓		✓	✓	✓	✓	✓	✓	• Entire band has co-primary use • Heavy Federal/DoD Allocation • Most EU states ambivalent at best • Passive (EES) Satellite & Radio Astronomy co-existence • ITU Decisions WRC-27 or later	• Co-existence/Sharing • Coverage and Link Budget vs. Cell Density
16-24GHz						✓			✓			• "FR2-like" (more challenging than <16GHz)
24-52 GHz			✓	✓		✓	✓	✓	✓	✓	• 24-52 Allocated allocated to Mobile IMT use	• Coverage • Energy Efficiency • Mobility
52-71GHz			✓	✓			✓	✓	✓	✓	• 57-71 Unlicensed	
71-110GHz		✓	✓			✓		✓	✓	✓	• Point-To-Point (71-76/81-86) & Automotive Radar • Inadequate contiguous sub-bands. • Heavy constraints 90-110	• Coverage • Energy Efficiency • Noise BW • Mobility
110-170GHz		✓	✓		✓	✓	✓	✓	✓	✓	• Lightly regulated • ITU RR-5.340 Constraints: Radio Astronomy/EES • ITU decisions WRC-31 or later	• Coverage • Energy Efficiency • Link Budget • Noise BW • Mobility
>170GHz		✓					✓	✓	✓	✓	• Lightly regulated so far • ITU RR-5.340 Constraints: Radio Astronomy/EES • ITU Decisions WRC-31 or later	

6G New Spectrum

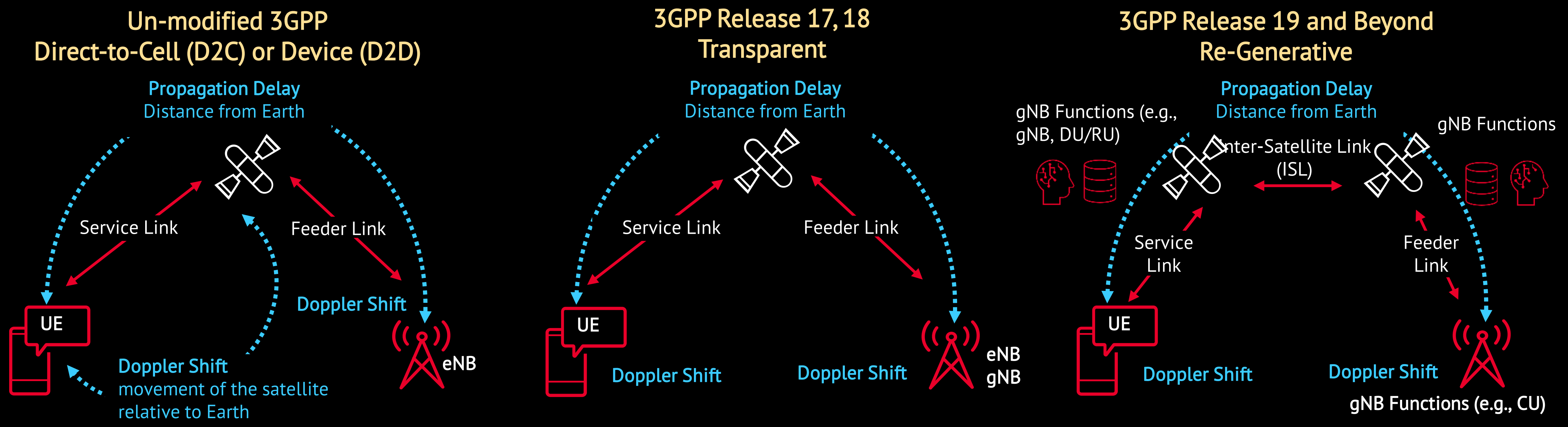
"FR3" and sub-THz under Evaluation



Satellite Communications – 3GPP NTN Architectures

Space to Earth – Satellite Options

Common Challenges: round-trip delays and frequency shifts due to the movement of the satellite relative to Earth (doppler shift)



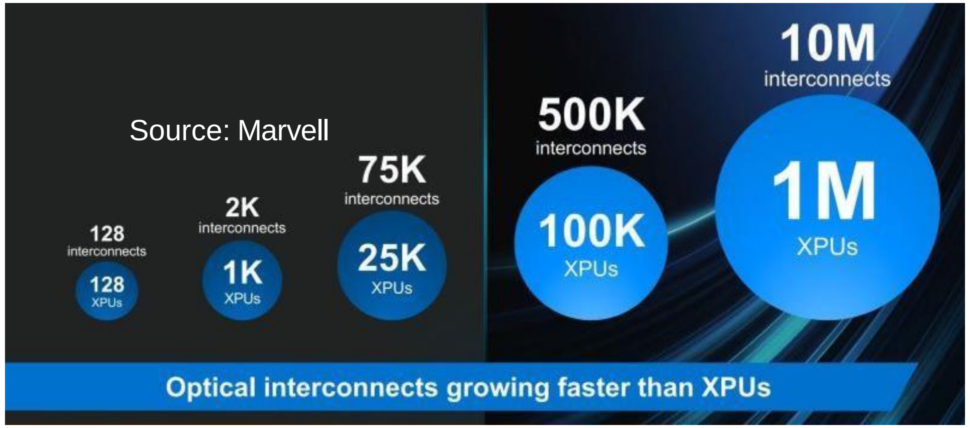
5G Advanced leading to 6G– Smarter with AI/ML

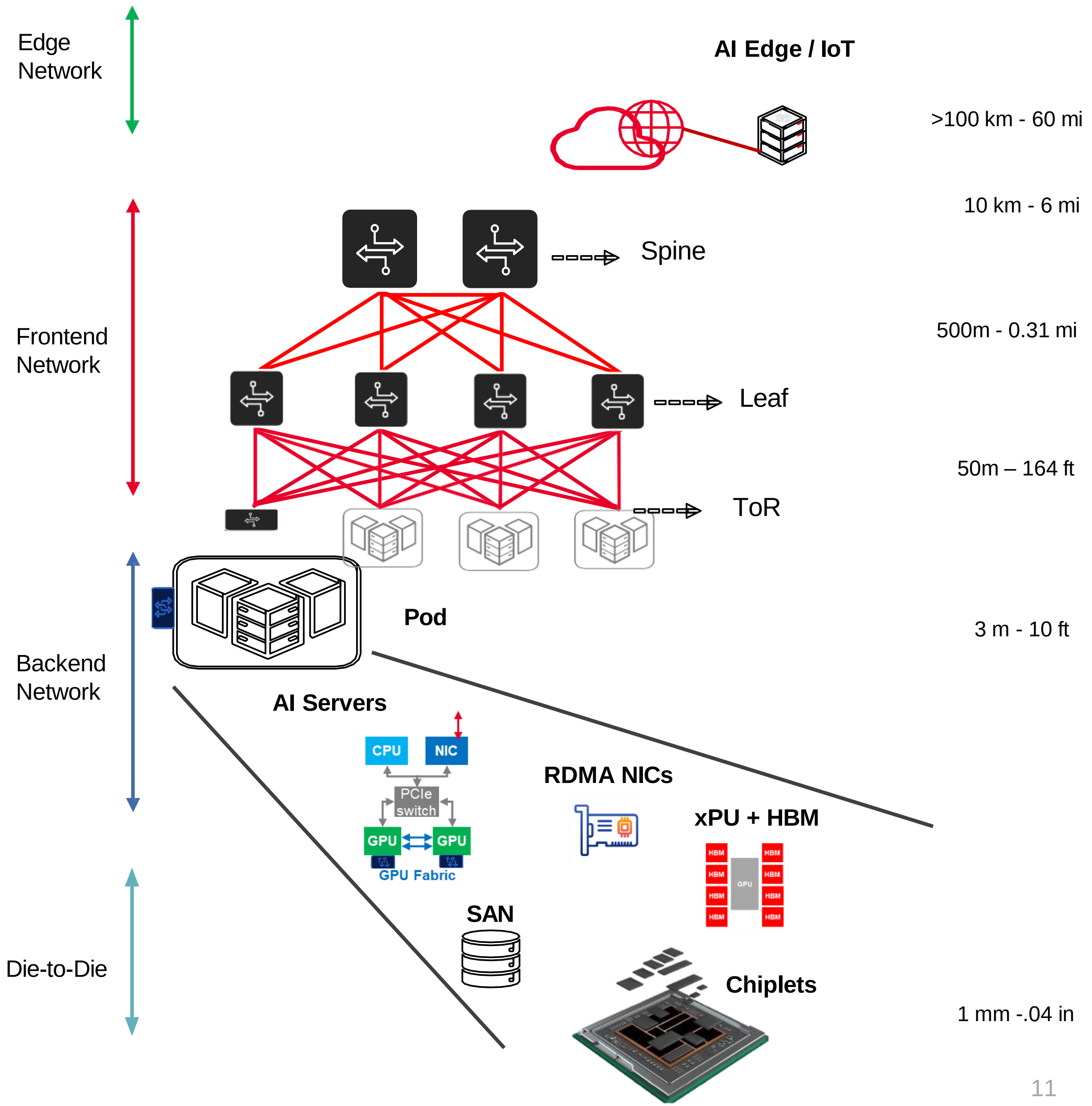
	Rel-17	Rel-18	Rel-19
NR Air Interface		CSI enhancements Beam Management Positioning enhancements	CSI enhancements Beam Management Positioning enhancements
Mobility			AI based Prediction: HO failure/RLF Measurement events Cell-level meas.
NG-RAN	Load Balancing Mobility Optimization Network Energy Optimization	Load Balancing Mobility Optimization Network Energy Optimization	Coverage & Capacity Opt Network Slicing

AI Infrastructure

Adapting Hyperscale DC to Edge AI

- Training Clusters: 100k+ GPUs in 2024 and **path to 600k**
- 800G/1.6T links, 112/224G lanes and path to 448G
- Power need **100+MW, 160% increase by 2030**
- New protocols for transport and congestion management



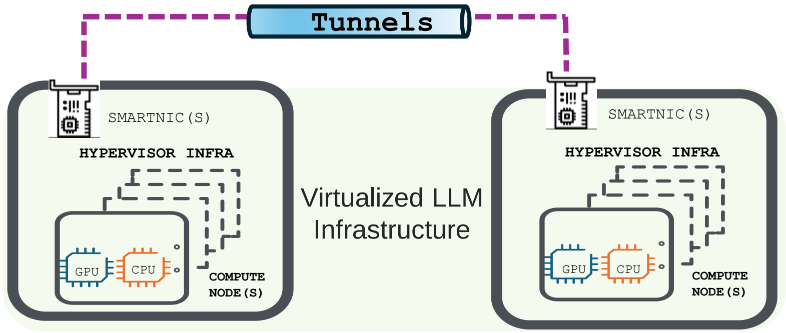


The Operation of AI ML Network Infrastructure

Backend Data Center for AI Models Training

East-West Traffic Test Demands -

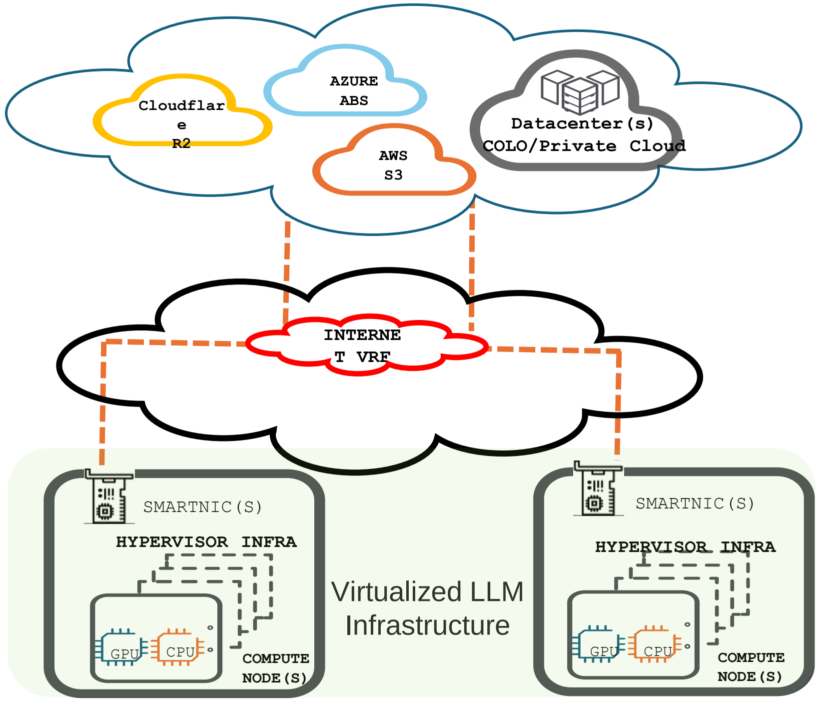
- Distributed GPU/CPU architectures
- Collective communications & parallel processing among GPU nodes
- Hyper-virtualized infrastructures for multi tenancy
- Immense performance needs for lossless connectivity and minimum tail-end latency



Front-end Data Center for Inference Workloads

North-South Network Traffic Test Demands -

- GPUs need high-speed access to block/remote storages
- Provisions to secure data in motion
- Ultra-low latency demands

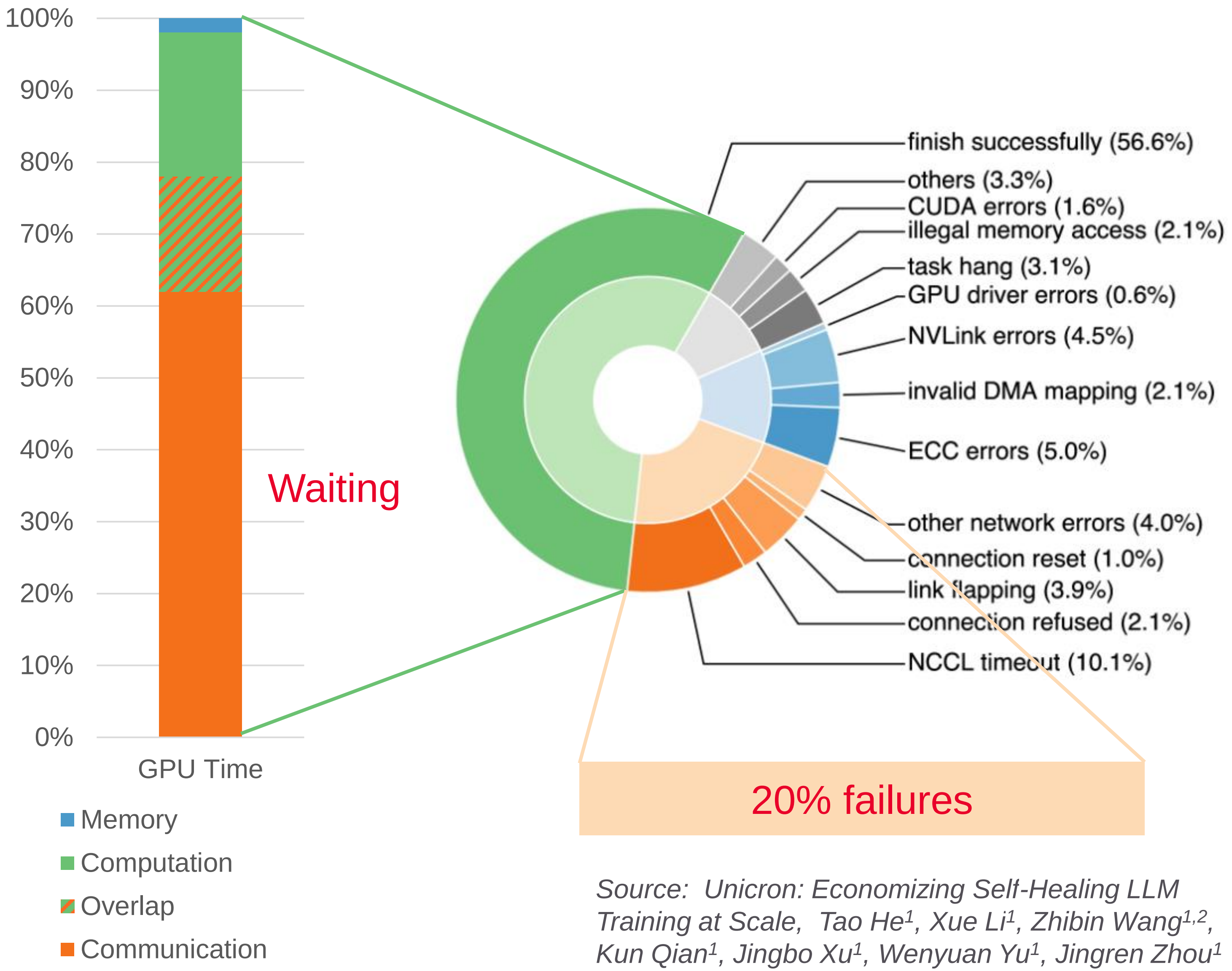


Why the Network & Components Matters in an AI Cluster

AI is Compute, Network & Data Intensive and requires validation at System Scale

Network failures
>20%

GPUs waiting on data
>50%



Source: Unicron: Economizing Self-Healing LLM Training at Scale, Tao He[1], Xue Li[1], Zhibin Wang[1,2], Kun Qian[1], Jingbo Xu[1], Wenyuan Yu[1], Jingren Zhou[1] [1]Alibaba Group, [2]Nanjing University

Vision transformer example. Source: https://github.com/facebookresearch/HolisticTraceAnalysi

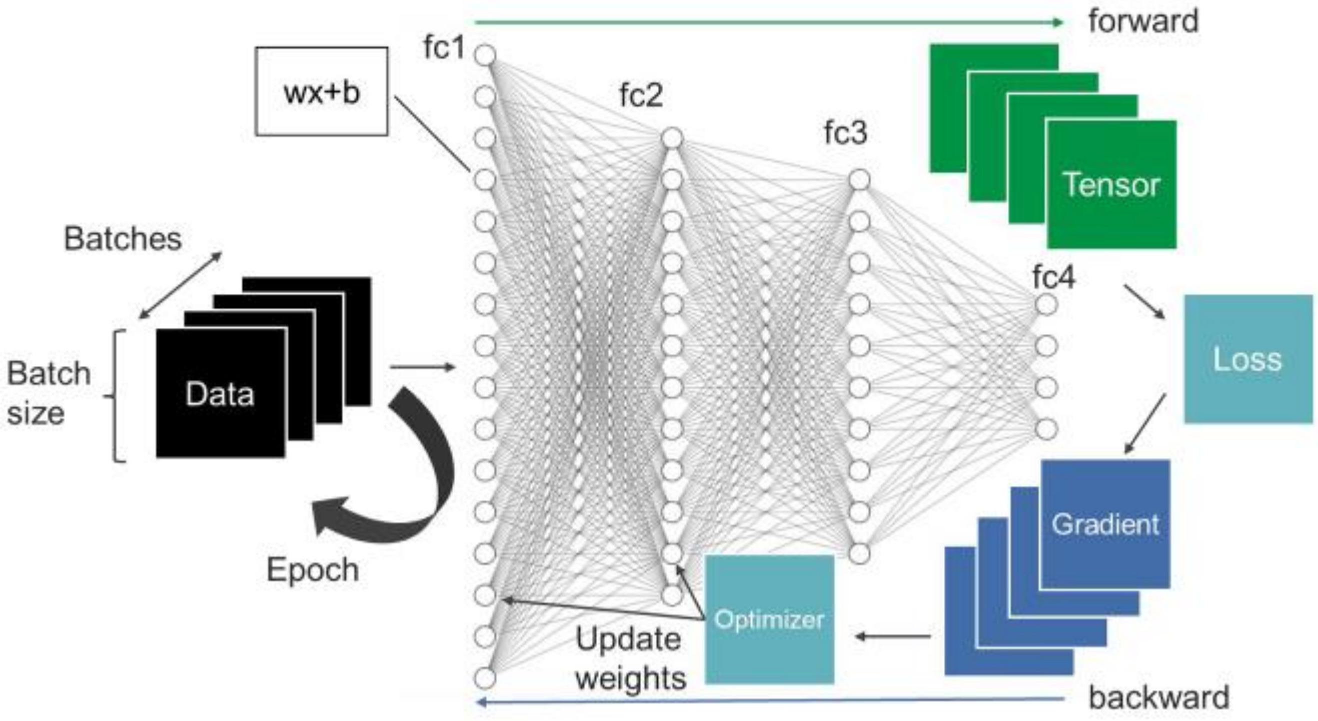
AI Model Training

3 Step Process

Step 1: Data preparation
- Collect and preprocess large datasets (for example, text files, images, and audio).
- Tokenize and normalize data to ensure consistency and efficiency.
- Split data into training, validation, and testing sets.

Step 2: Model definition
- Define the architecture of the AI model (for example, neural network and decision tree).
- Specify hyperparameters (for example, learning rate, batch size, and number of layers).

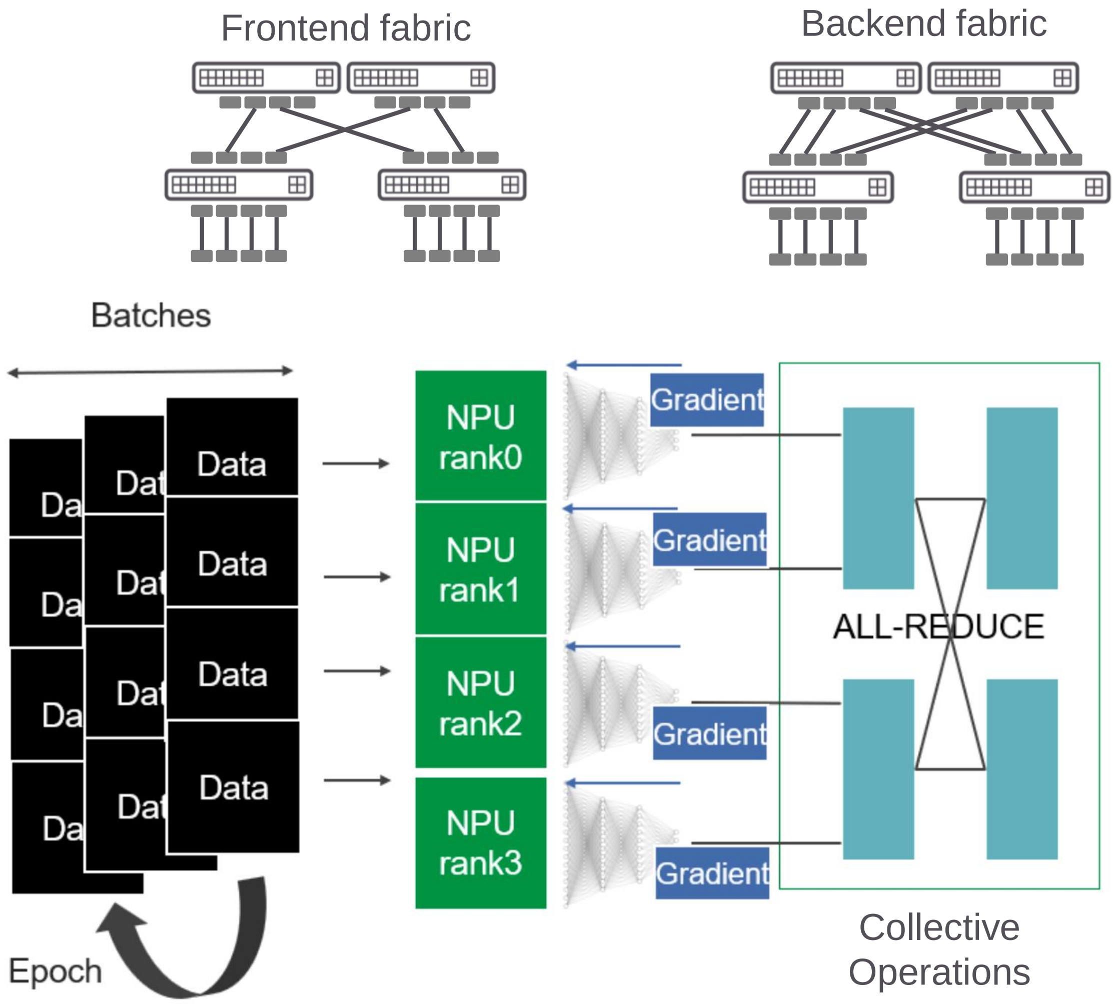
Step 3: Model training
- Initialize the model's weights and biases.
- Feedforward pass: Compute outputs for each sample in the training set.
- Backpropagation: Calculate gradients and update model parameters by using an optimization algorithm (for example, Stochastic Gradient Descent and Adam).
- Repeat the preceding steps until convergence or a stopping criterion is reached.



Network role in AI clusters

Scaling up systems, scaling out clusters

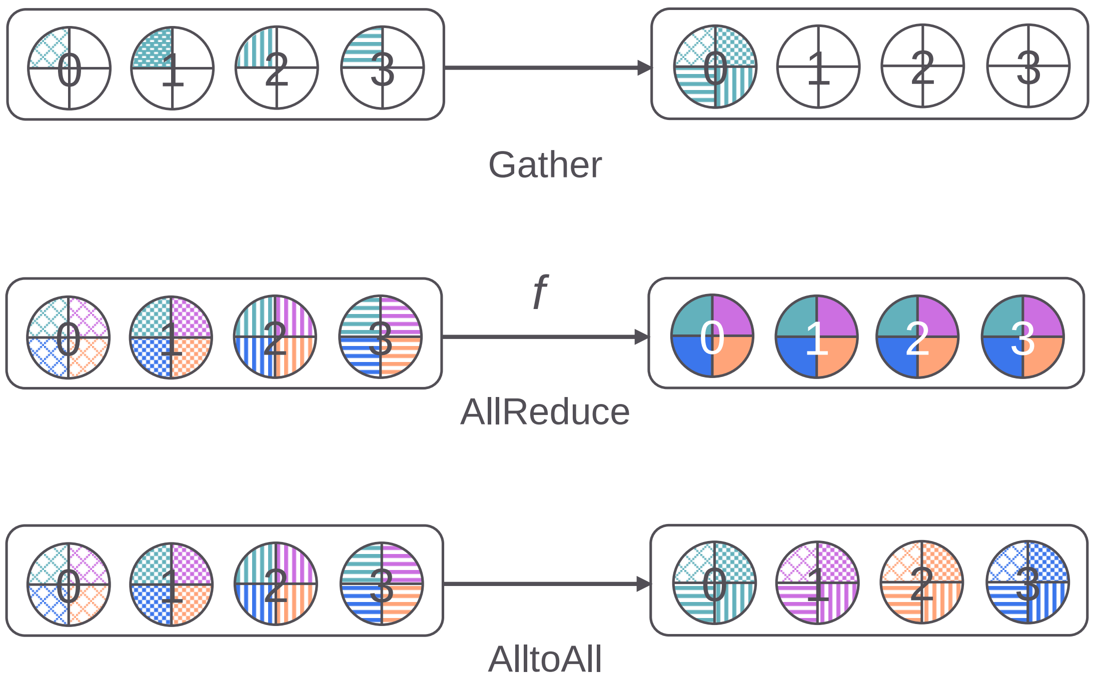
- Accelerate model training with **Data Parallelism**
- Split large models across GPUs with **Tensor** and **Pipeline Parallelism**
- Subdivide complex problems among several models with **Mixture of Experts**



Data Parallel Architecture

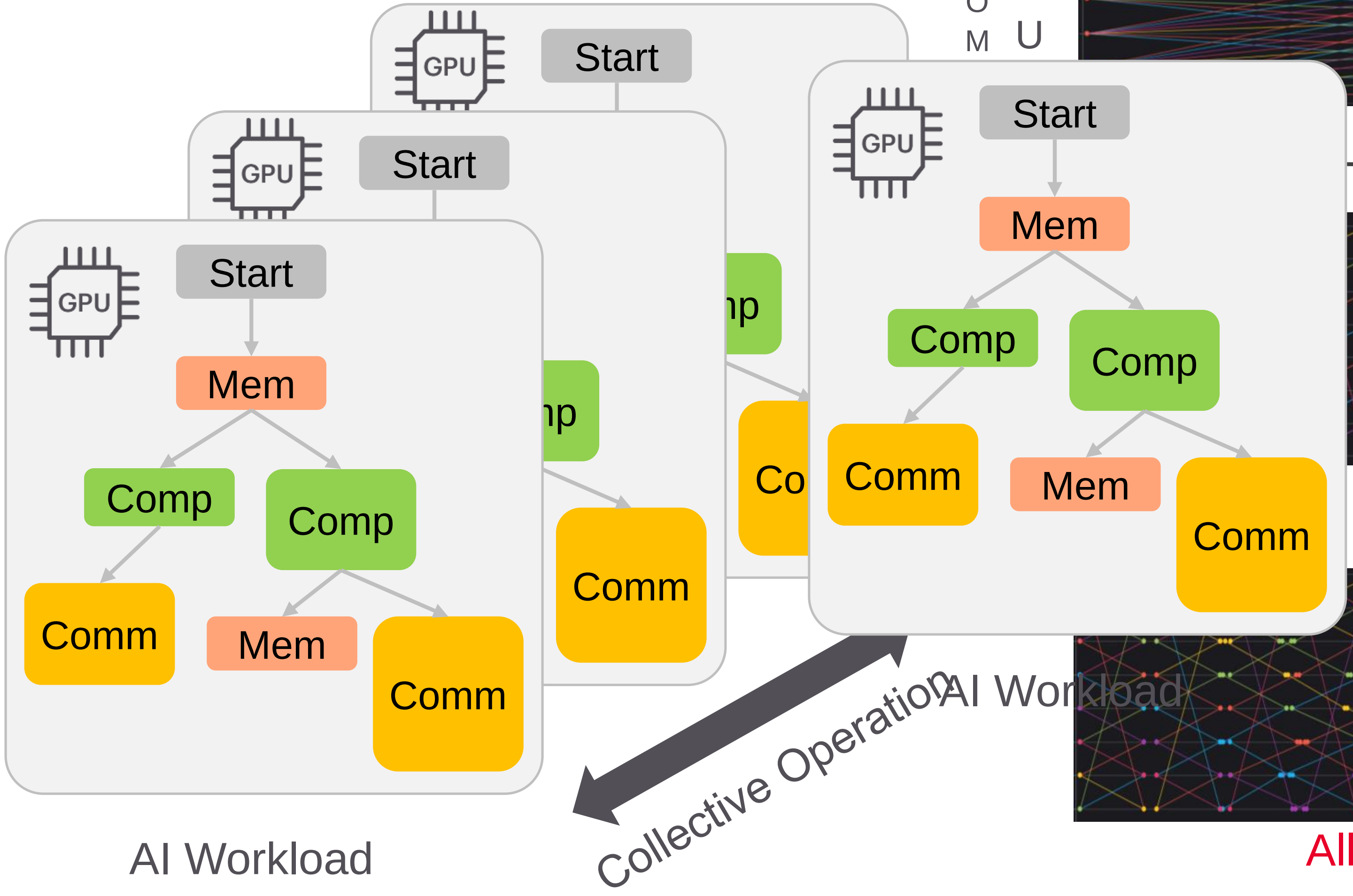
Types of Collective Operations

- Common types for AI workloads:
  - Broadcast
  - Gather
  - AllReduce
  - AlltoAll
  - ReduceScatter
  - AllGather
- Reduce implies math with data (*f*)
- *All* or *Scatter* – symmetry



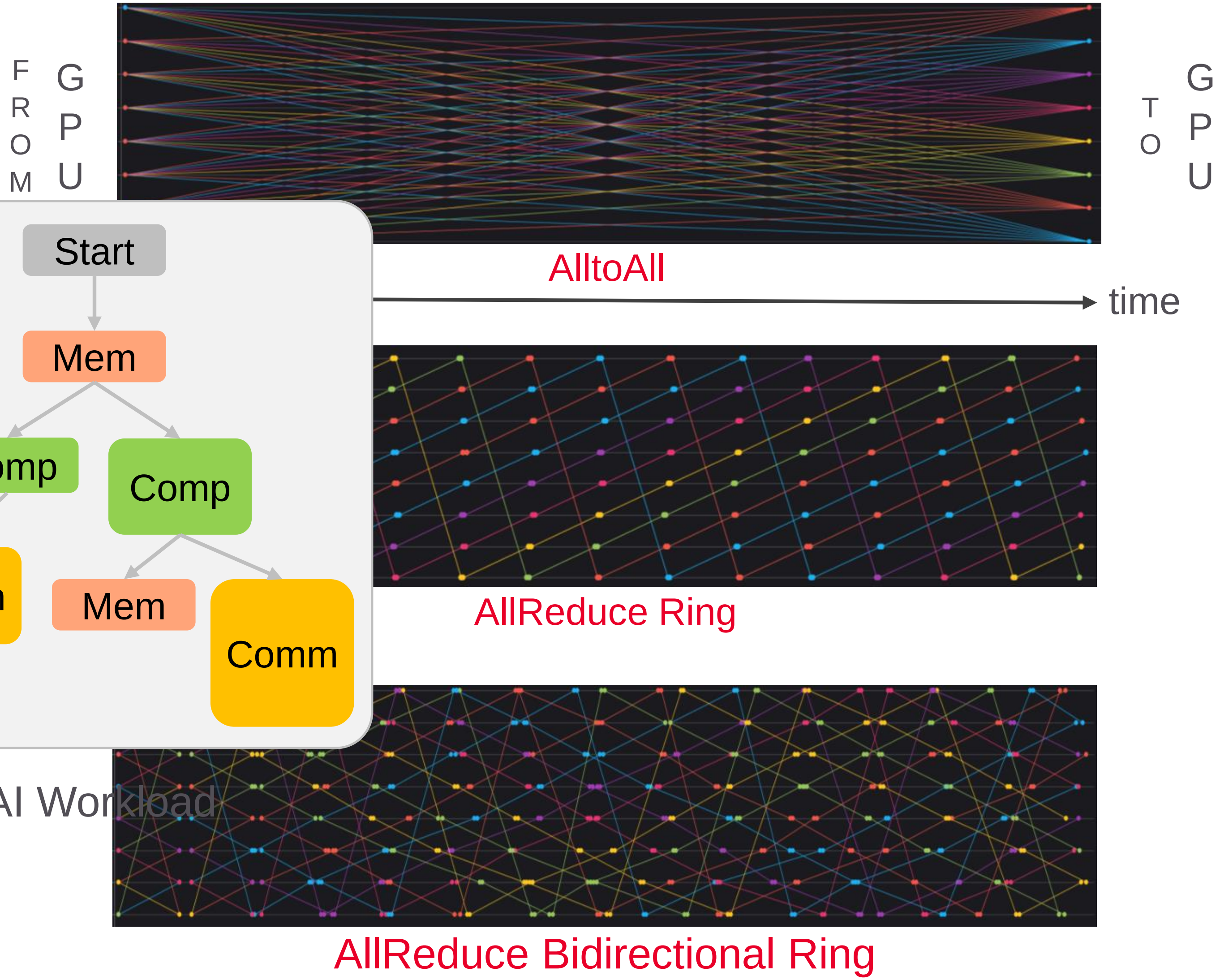
GPU Communications

Breaking down AI workload



AI Workload

Examples of Collective Operations



Network is the bottleneck in AI model training

Job Completion Time Factors

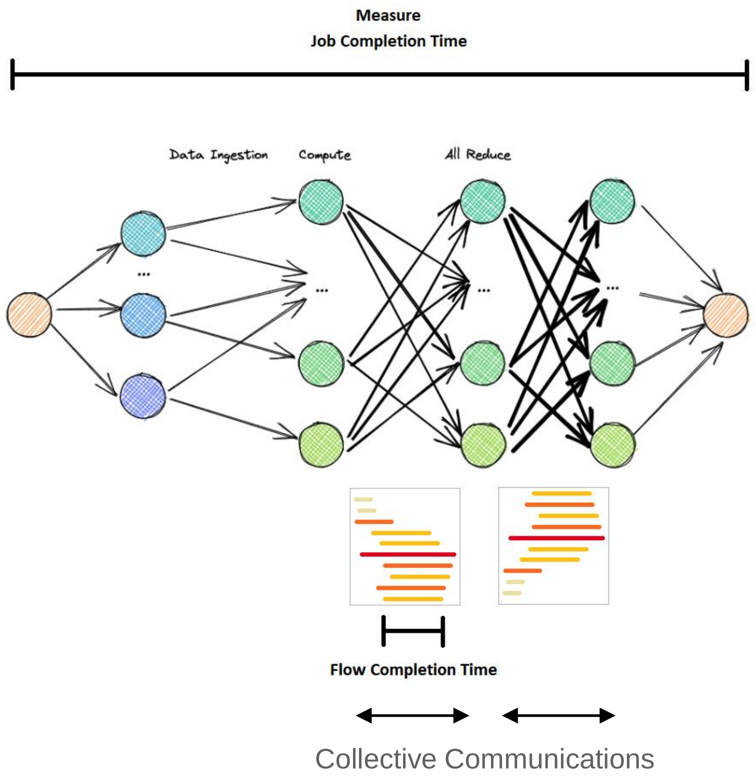
- Data Ingestion
- Computation
- Collective Communications

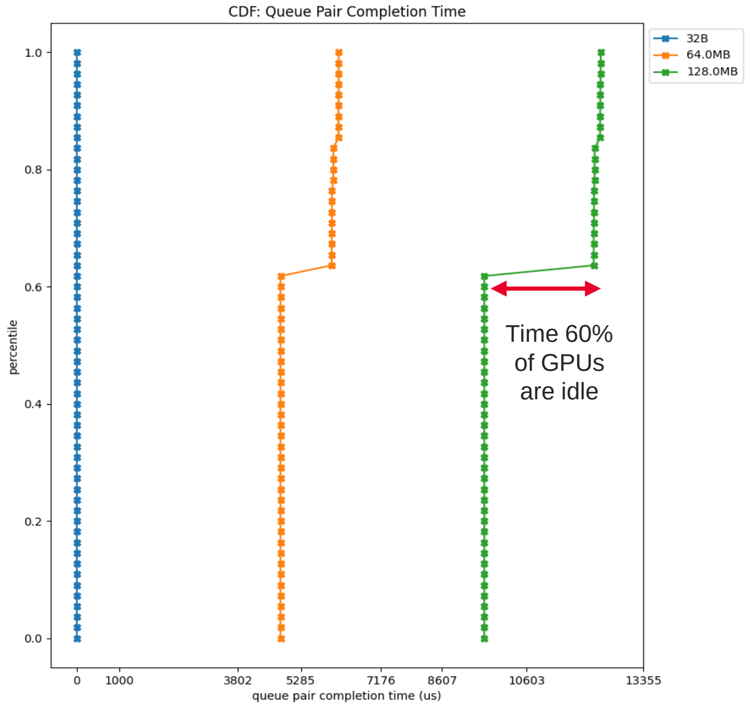
Network tail latency

- Defines wasted GPU time

Contributors

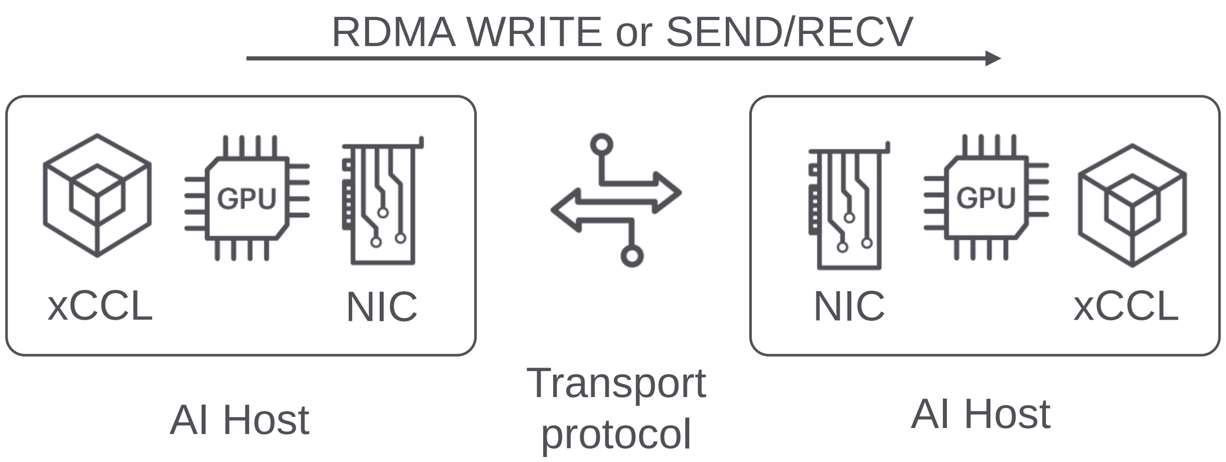
- Data exchange algorithm
- Software stack
- System I/O
- DPU (NIC)
- Network fabric





RDMA and Transport Protocols

Hardware accelerated Remote Direct Memory Access

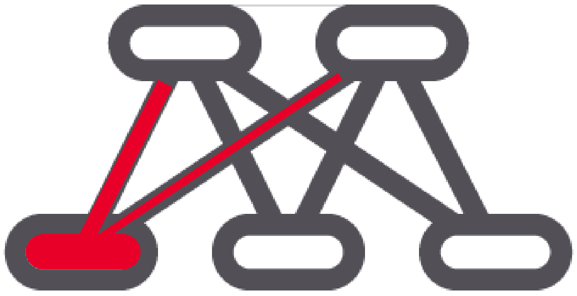


Ethernet transport options

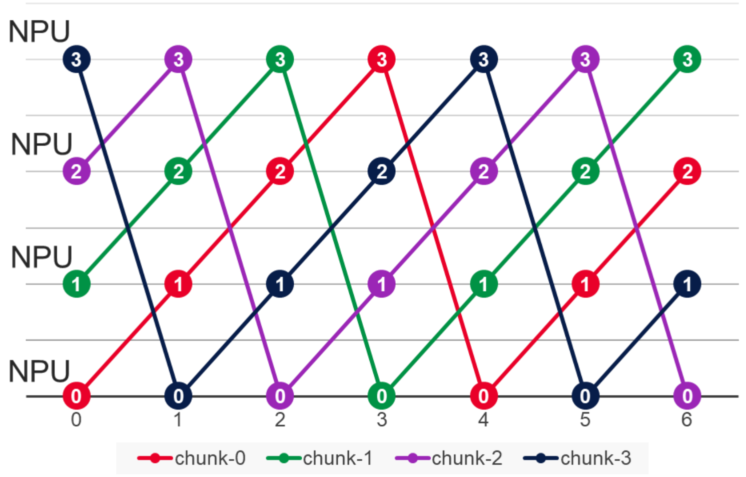
- RoCEv2
- Falcon
- Custom / Proprietary
- Ultra Ethernet (future)

Practical challenges

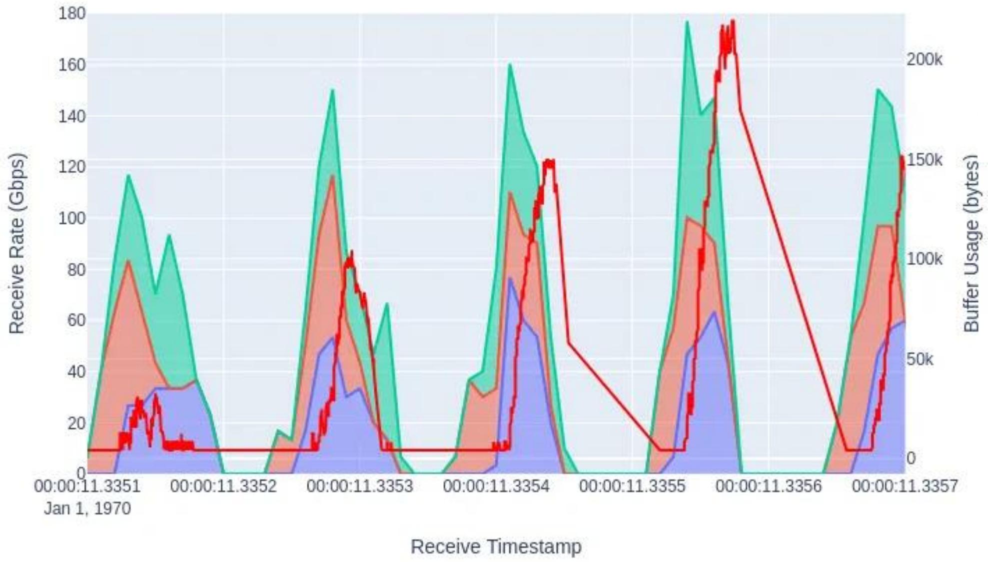
- Synchronized start – from 0 to line rate on all ports
- Flow dependencies – latencies accumulate
- Low entropy – hard to load balance
- RDMA message bursts – incast



Unequal Load Balancing

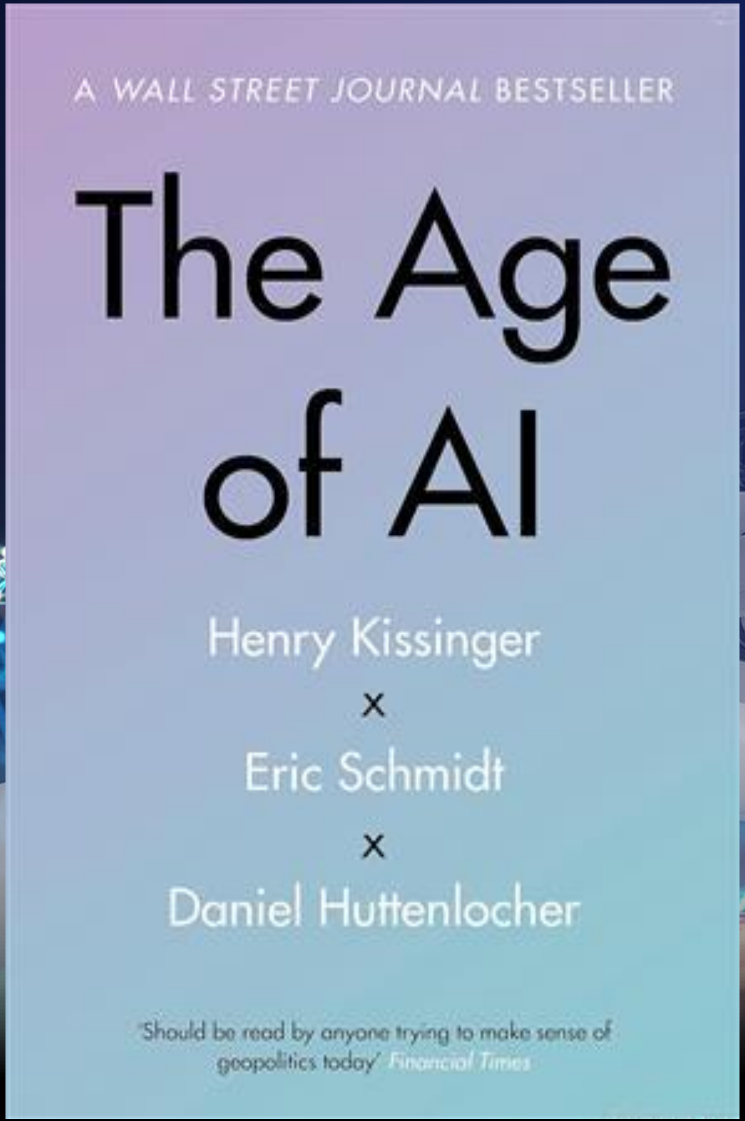


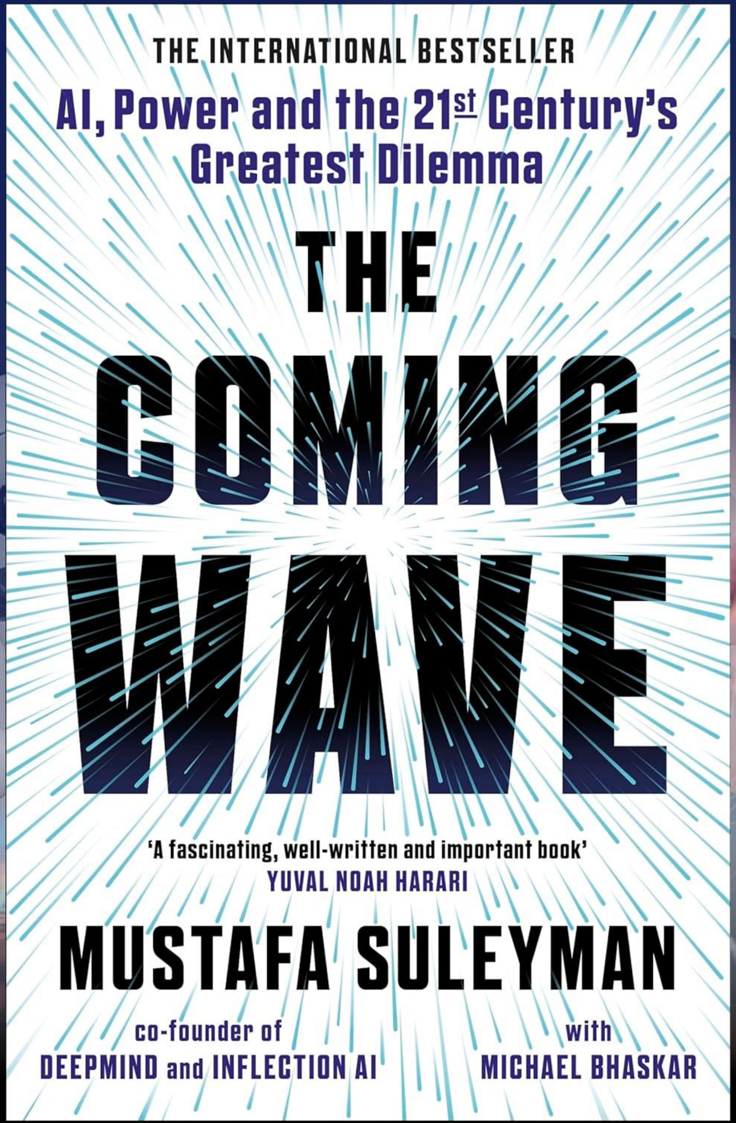
Latencies

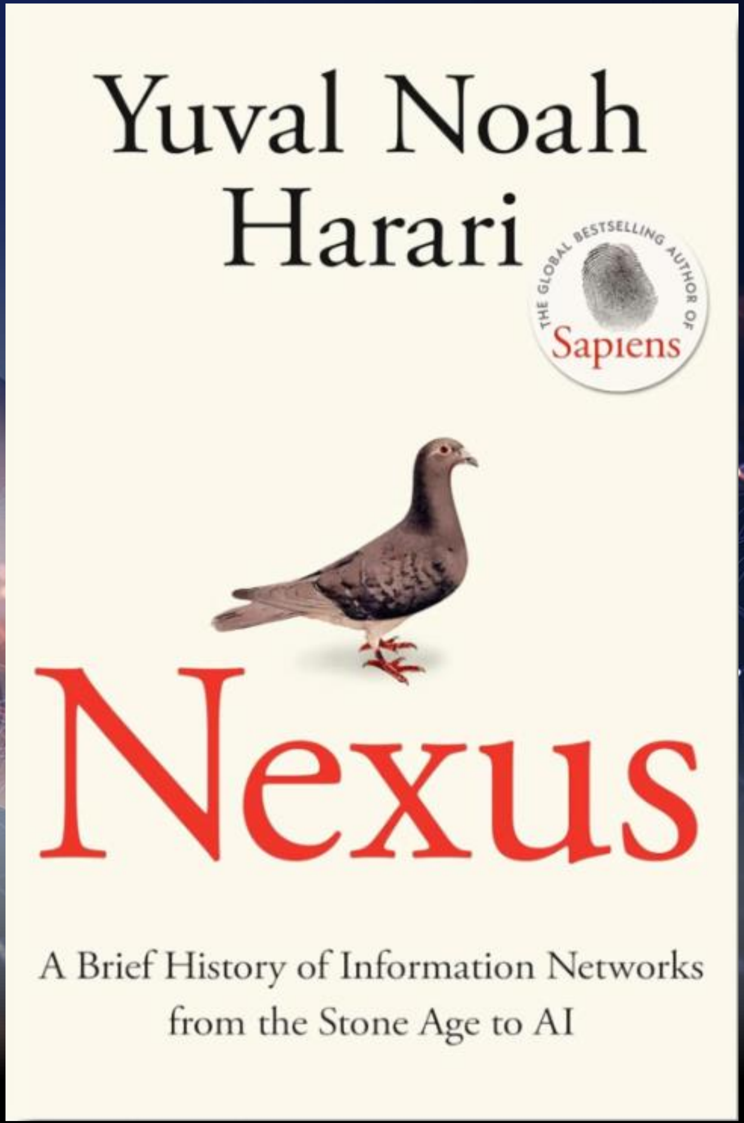


Incast

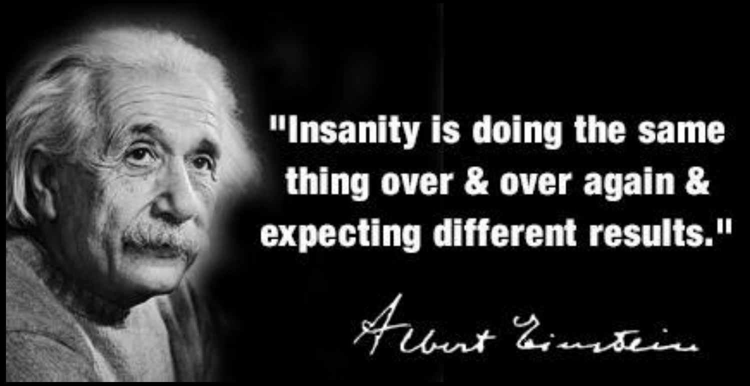
Suggested Reading………

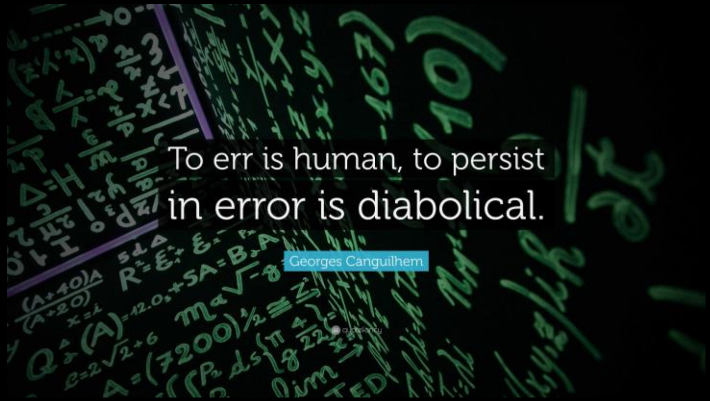


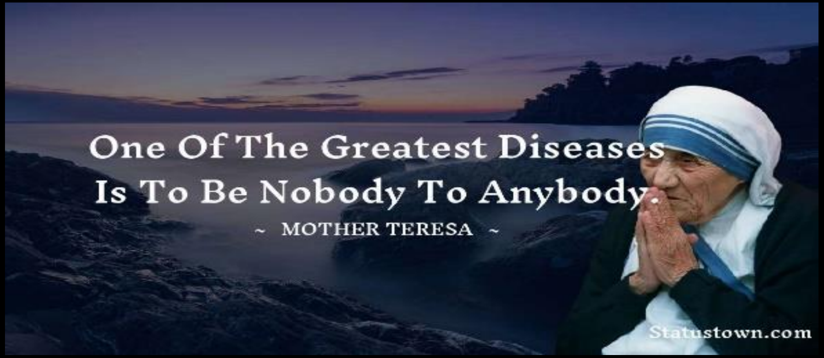


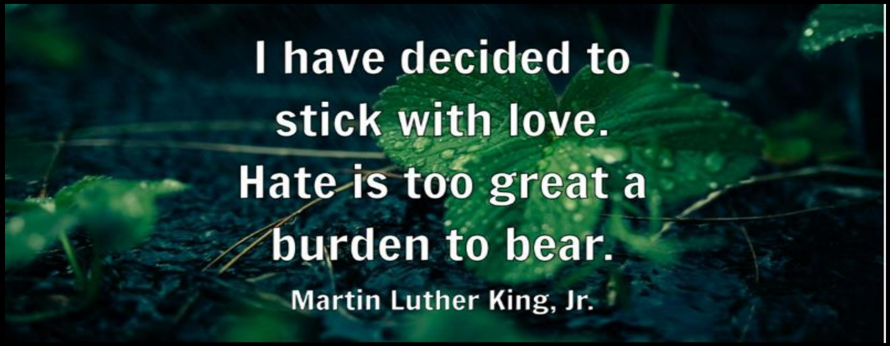


Finally, Some of my favorite quotes









"The good thing about science is that it's true whether you believe in it or not"

Neil deGrasse Tyson

KEYSIGHT