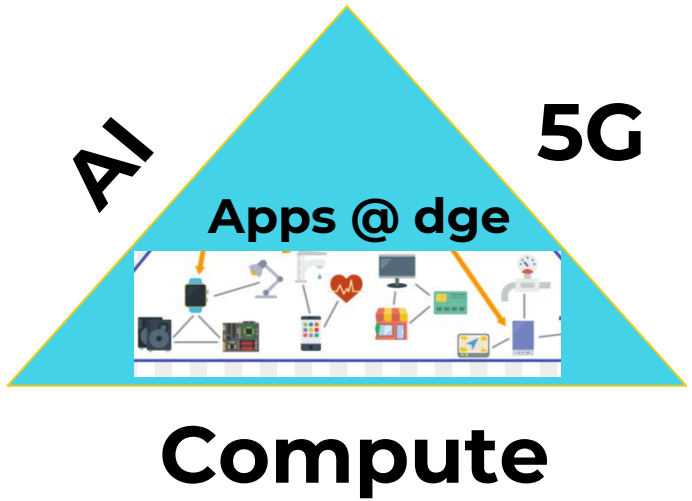
GPU TECHNOLOGY CONFERENCE

Reinventing Edge Computing Applications by harnessing the power of AI, GPU, & 5G

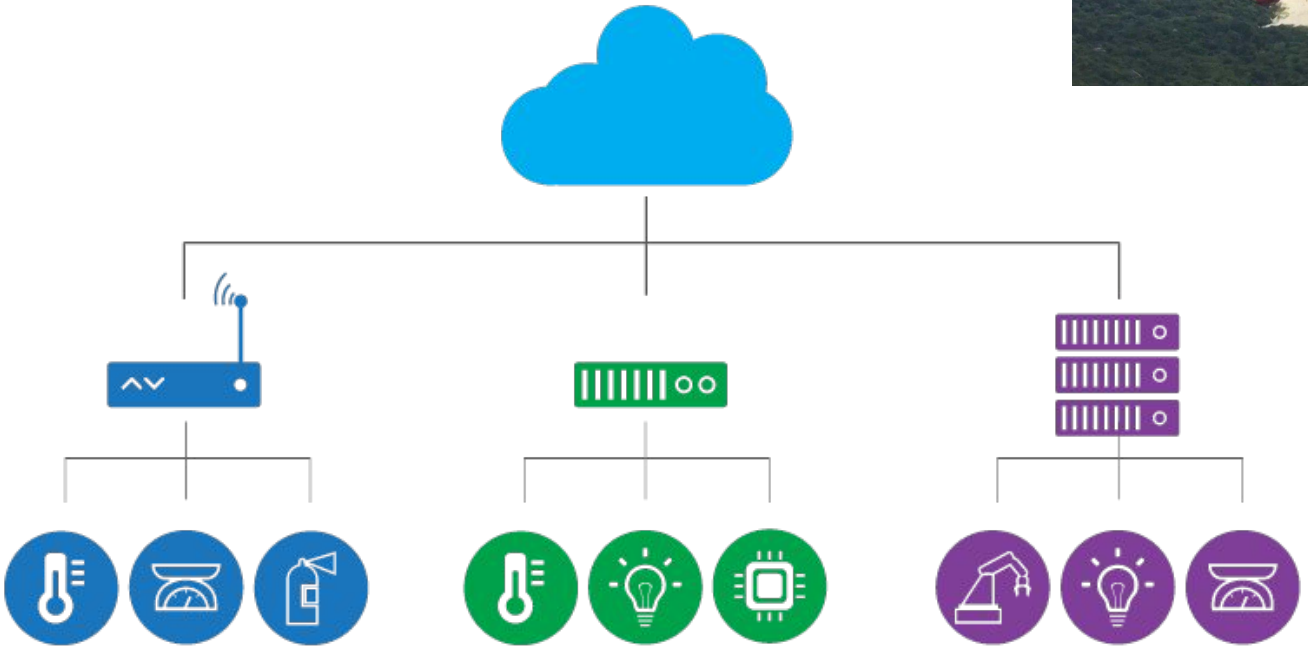


Gyana Dash
gyana.dash@gmail.com

Goal of this Session

- **Edge Computing Ecosystem**
  - **What it is & Why we need**
- **Brief on AI, GPU & 5G**
  - **How each accelerates Edge Computing**
- **Rethink Edge Computing Applications**
  - **Use cases focusing on Humanity & Environment**
- **Case Studies**
  - **Project and Research Papers**

Where is the Edge



1. Camera, Mobile Phone, Sensors, Drones, Robots
2. Cell Towers, Gateways, Wi-Fi Access Points
3. Data Centers
4. Telecom Core, Internet
5. Cloud Services

What is Edge Computing

- Gartner defines edge computing as solutions that facilitate data processing at or near the source of data generation.

- Edge computing promises near real-time insights and facilitates localized actions.

- **Example:** The AWS DeepLens camera integrates a 1080p camera with a Linux operating system and specialized ML software integrated into a business application.

Why Edge Computing

- Speed, Time and Latency
- Cost and Volume of Data Transfer
- Privacy and Regulatory Compliance

Smart Cities

Why Edge Computing

- Speed, Time and Latency
- Cost and Volume of Data Transfer
- Privacy and Regulatory Compliance

Industry & Extreme Condition

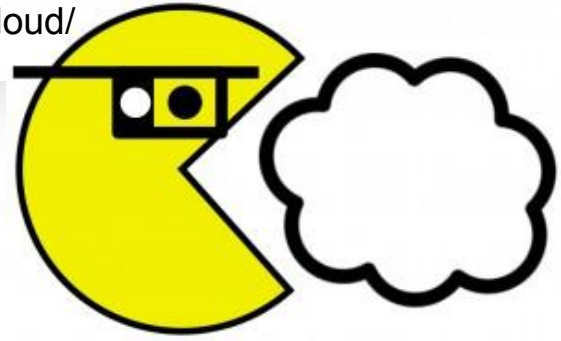
Why Edge Computing

- Speed, Time and Latency
- Cost and Volume of Data Transfer
- Privacy and Regulatory Compliance

Medical Equipment

Edge Computing - Gartner Strategy

https://blogs.gartner.com/thomas_bittman/2017/03/06/the-edge-will-eat-the-cloud/



[The Edge](#)

[Will Eat The Cloud](#)

AI Foundations
Intelligent Apps & Analytics

Smart Devices
Edge Computing
Cloud Computing

Collaborative
AR/VR
Blockchain

Brief on AI

Help in Cleaning

OR

Call 911 for Help

DeepLens

Brief on AI

- **Lot of Data for human to process & Inference**
  - **Unsupervised - text, categorical data**
  - **Supervised - text, picture, video**
- **Automate Tasks to improve productivity**
  - **Supervised/Semi-supervised**
- **Robotics to improve human life**
  - **Learning by doing**
  - **Learning by Observing**

Brief on 5G Network Functions

Source: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.04.00_60/ts_123501v150400p.pdf

1. Authentication Server Function (AUSF)
2. Access & Mobility Management Function (AMF)
3. Data Network (DN), e.g. operator services, Internet access or 3rd party services
4. Unstructured Data Storage Function (UDSF)
5. Network Exposure Function (NEF)
6. Network Repository Function (NRF)
7. Network Slice Selection Function (NSSF)
8. Policy Control Function (PCF)
9. Session Management Function (SMF)
10. Unified Data Management (UDM)
11. Unified Data Repository (UDR)
12. User Plane Function (UPF)
13. Application Function (AF)
14. User Equipment (UE)
15. (Radio) Access Network ((R)AN)
16. 5G-Equipment Identity Register (5G-EIR)
17. **Security Edge Protection Proxy (SEPP)**
18. Network Data Analytics Function (NWDAF)

Brief on 5G System Architecture

Network

Session

User

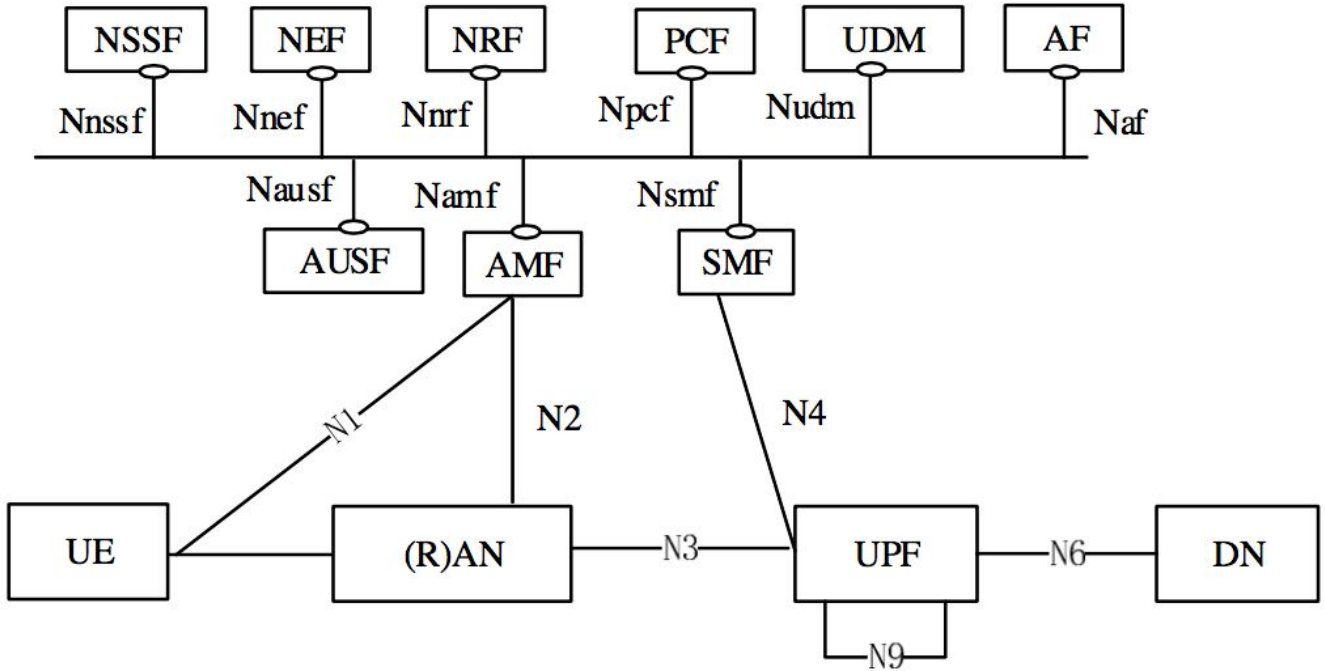


Figure 4.2.3-1: 5G System architecture

Source: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.04.00_60/ts_123501v150400p.pdf

Brief on 5G Data Storage

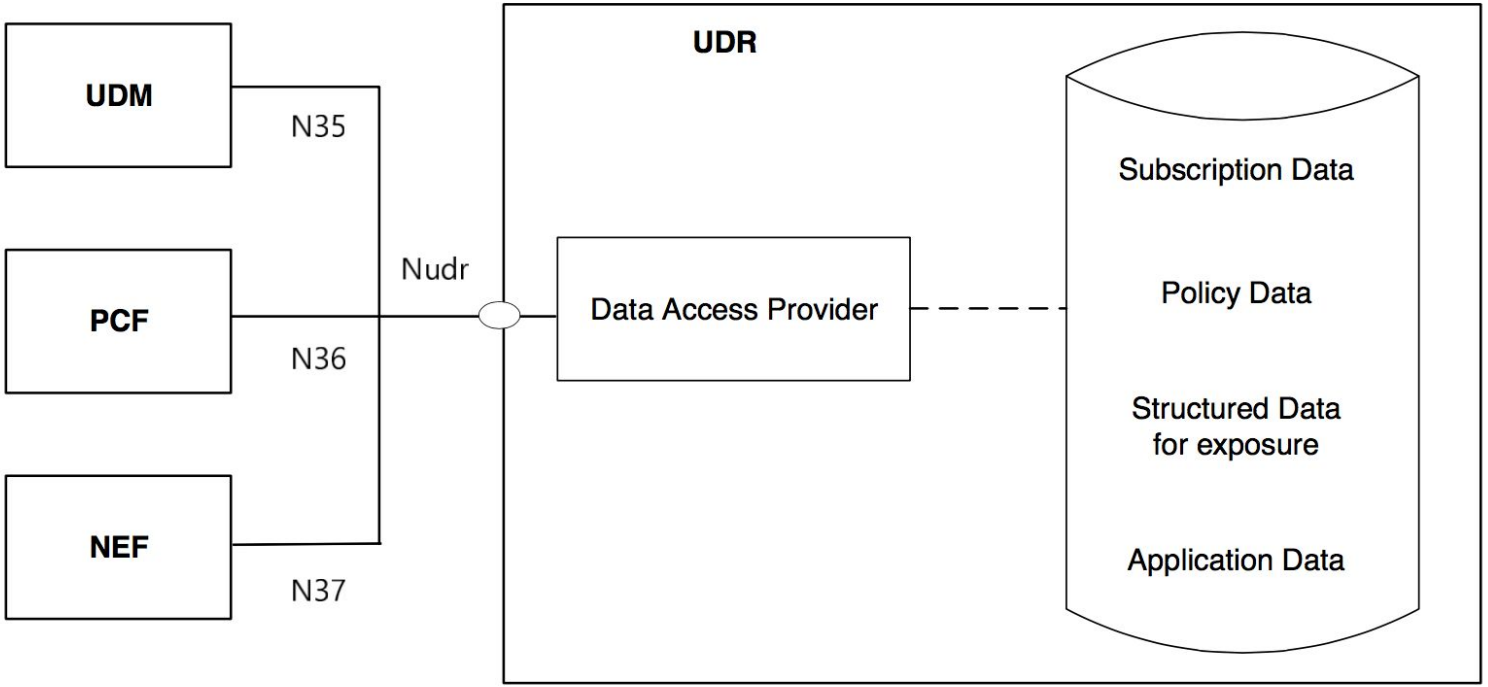


Figure 4.2.5-2: Data storage architecture

Source: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.04.00_60/ts_123501v150400p.pdf

5G Promises

CAPABILITY	5G TARGET	USAGE
Peak Data Rate	20 Gbps	eMBB
User Experienced Data Rate	**100 Mbps - 1 Gbps**	eMBB
Latency	**1 ms**	URLLC
Mobility	500 km/hr	eMBB/URLLC
Connection Density	10^6 /km^2	MMTC
Energy Efficiency	Equal to 4G	eMBB
Spectrum Efficiency (BW throughput)	**3 - 4X of 4G**	eMBB
Area Traffic Capacity	1000 (Mbit/s)/m^2	eMBB

eMBB - enhanced Mobile Broadband or handsets

URLLC - Ultra-Reliable Low-Latency Communications or autonomous

MMTC - Massive Machine Type Communications or sensors

Source: https://en.wikipedia.org/wiki/5G

5G Network Platforms

Brief on GPU

You all agree to Skip Right?

Yes, we are in GTC

Edge Computing Applications

AI

Sector	Subsector	Potential AI Consumption Impact	Personalisation	Time Saved	Utility	Data Availability
Healthcare		3.7	3.8	2.7	3.9	4.4
	Providers/Health Services	3.9	4.1	3.0	3.9	4.7
	Pharma/Life Sciences	3.8	3.9	2.8	4.2	4.1
	Insurance	3.6	3.6	2.6	3.8	4.2
	Consumer Health	3.5	3.4	2.3	3.4	4.8
Automotive		3.7	3.9	2.9	3.8	3.9
	Aftermarket & Repair	3.9	4.2	2.8	3.6	4.6
	Component suppliers	3.9	4.0	2.0	3.5	5.0
	Personal Mobility as a Service	3.8	4.0	3.7	4.0	3.7
	OEM	3.6	4.0	3.0	4.0	3.5
	Financing	3.3	3.3	3.0	3.7	3.0
Financial Services		3.3	2.8	2.6	3.2	4.6
	Asset Wealth Management	3.4	2.9	2.2	3.7	4.3
	Banking and Capital	3.3	2.5	2.9	3.0	5.0
	Insurance	3.2	3.1	2.4	3.1	4.1
Transportation and Logistics		3.2	3.5	2.6	3.3	3.7
	Transportation	3.5	3.0	2.8	3.5	3.0
	Logistics	3.1	3.9	2.5	3.1	3.0
Technology, Communications and Entertainment		3.1	2.5	2.1	3.3	4.3
	Technology	3.3	2.7	1.9	3.6	4.1
	Entertainment, Media and Communication	3.0	2.5	2.0	3.3	4.4
Retail		3.0	2.8	2.1	3.3	3.8
	Consumer Products	3.1	3.0	2.3	3.3	3.8
	Retail	2.9	2.6	2.0	3.3	3.7
Energy		2.7	3.2	2.1	3.1	3.1
	Oil & Gas	2.3	4.0	2.1	2.9	3.0
	Power & Utilities	2.1	2.0	2.1	3.3	3.2
Manufacturing		2.2	2.0	1.2	3.7	3.8
	Industrial manufacturing	2.2	2.0	1.4	3.7	3.9
	Industrial Products/Raw Materials	2.1	NA	1.0	3.6	3.7
Grand Total		3.1	3.1	2.3	3.5	4.1

Across All Industries

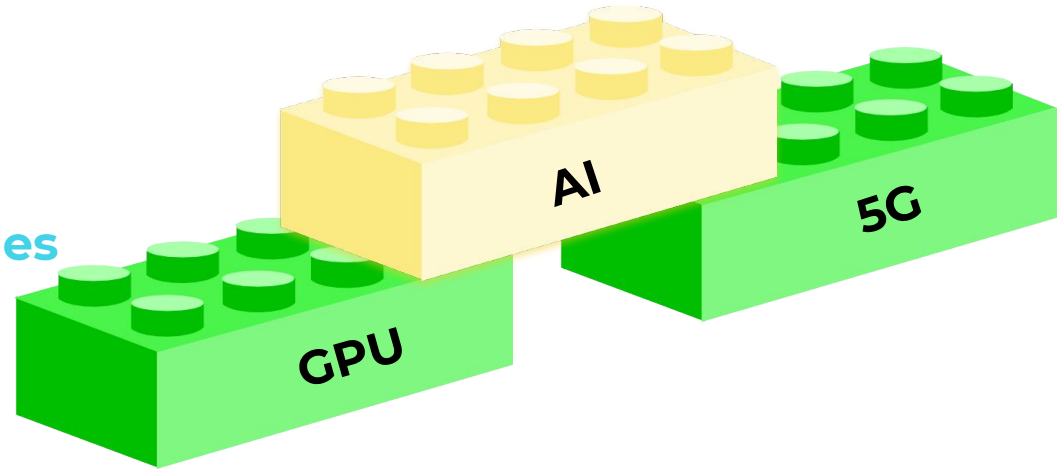
5G

	Challenge	Need	How 5G will help
Automotive	- Strict CO_2 emission goals - Strong competition - Pressure for innovation - Globalisation	- Autonomous and connected cars - Innovative infotainment solutions	- Dynamically configure networks and resources to address different demands
Media and entertainment	- Quality of experience constantly increasing - New devices and services - Explosion of mobile data usage	- Networks which can support new media and entertainment services and devices (VR & AR)	- Support massive increases in data rules - Guarantee a good quality of service
Energy and utilities	- Decentralised generation - Pressure on consumption - Increase in renewables - Fines when outage	- Dynamic smart grids, which can be monitored and controlled remotely throughout the entire network	- Real-time control of grids and remote generators where fibre has not been rolled out
Public transport	- Stronger focus on safety and security - Growing number of passengers - Higher service expectations	- Real-time information and entertainment for passengers - More efficient operations and maintenance of infrastructure	- Provide coverage and bandwidth for infotainment and more efficient operations
Agriculture	- Growing global population - Pressure on use of pesticides - Lack of farmers - Climate change	- Increased productivity and efficiency of farming - Sustainable farming solutions	- Remotely connect and control farming equipment - Provide bandwidth for advanced imagery and use of drones
Healthcare	- Ageing population - Increase in people with chronic diseases - Personalised care expectations	- Affordable healthcare solutions - Personal, wearable devices for monitoring and treatment - Remote patient care and follow up	- Enable mobile remote care solutions through guaranteed and secured connection
Manufacturing	- Ageing workforce - Manufacturing skills gap - Pressure on costs - More environmental concerns	- Robotics and automation inside the factory - Solutions which decrease production costs	- Provide the highly resilient, secure and low latency communication platform in the factory
Security	- Higher security alerts - Increased terrorist threats	- More monitoring and screening in public places - Better and faster	- Support wireless security applications both for

Source: Arthur D. Little, PWC, Gill Pratt

Leading Edge Computing Applications

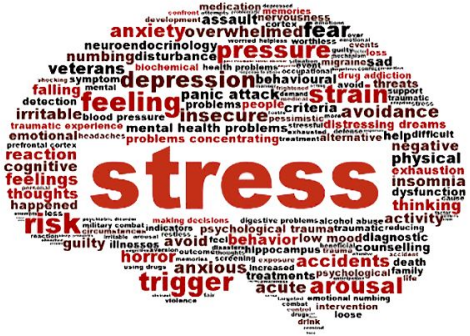
- **Immersive Experience**
  - AR/VR and Mixed Reality
- **Automotive & Transportation**
  - Connected Vehicles
- **Remote Operation**
  - Factory, Hospital
- **Intelligent Automation**
  - Industrial and Smart Cities



Edge Computing Applications: Humanity

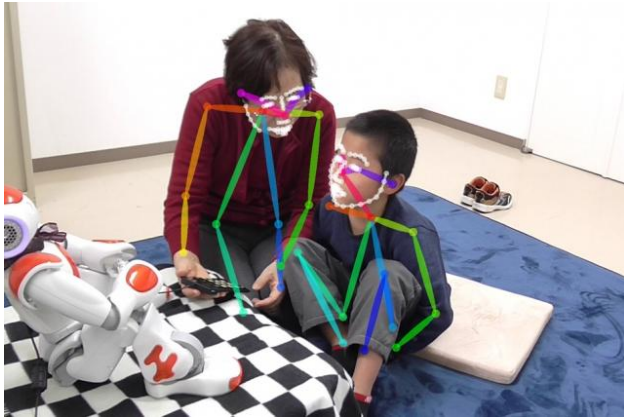
Alzheimer
5.8 M US
50M world
$290B US

Back Home Safely



Guide to manage
Stress

Autism:
Express the feeling



Edge Computing Applications: Environment

Collaborative Edge Computing Solutions to address Environmental issues and natural Disasters

Case Study 1 - Traffic Engineering - The Problem

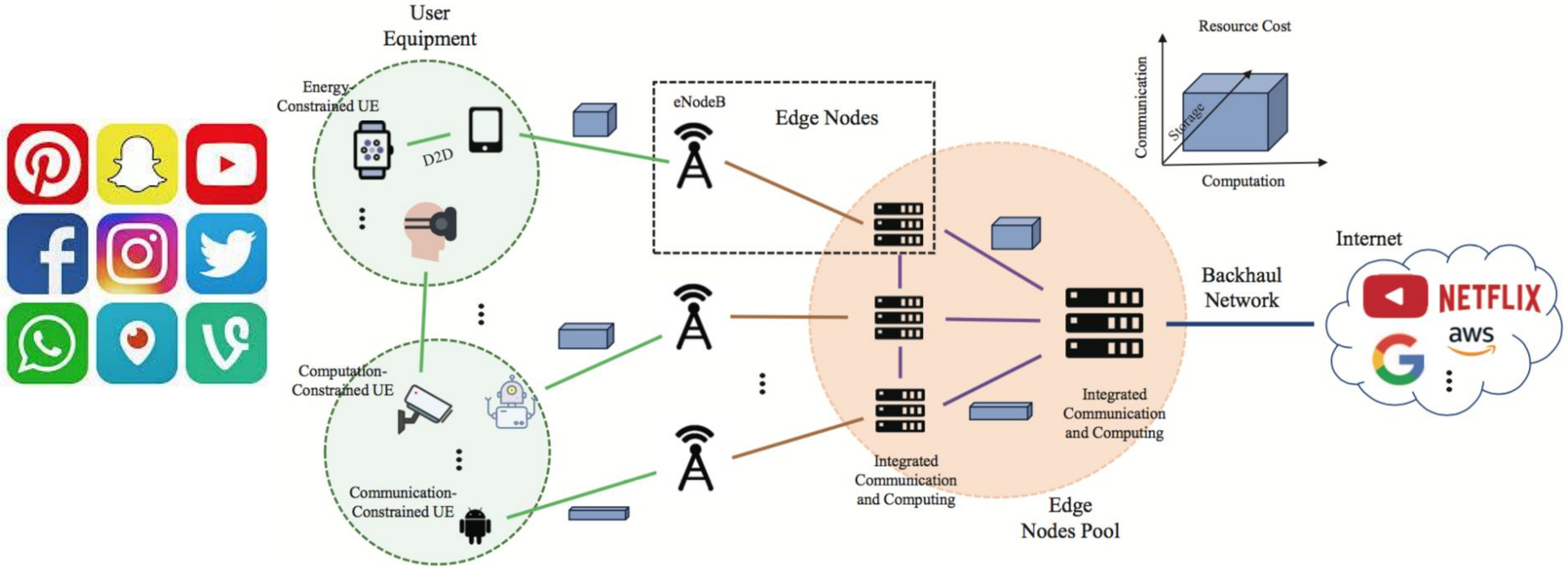
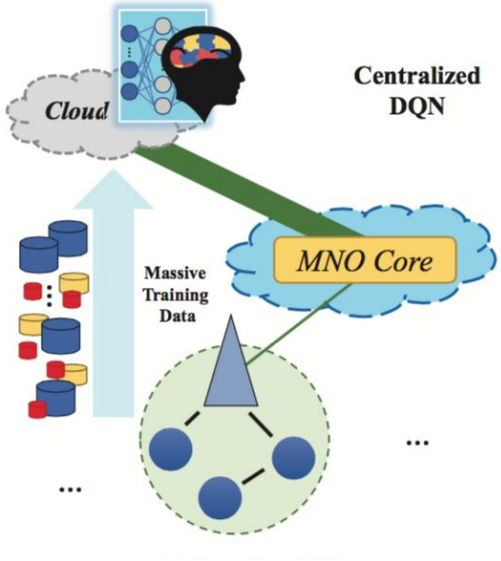


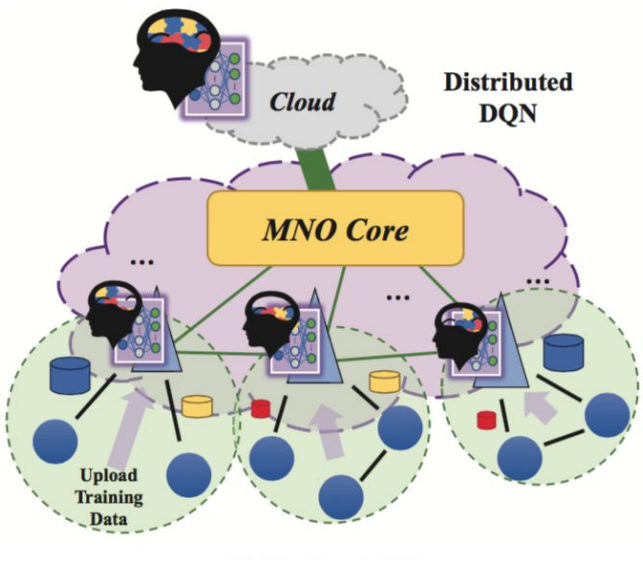
Fig. 1. Framework of AI-supported mobile edge system with cognitive ability

Source: https://arxiv.org/abs/1809.07857v1

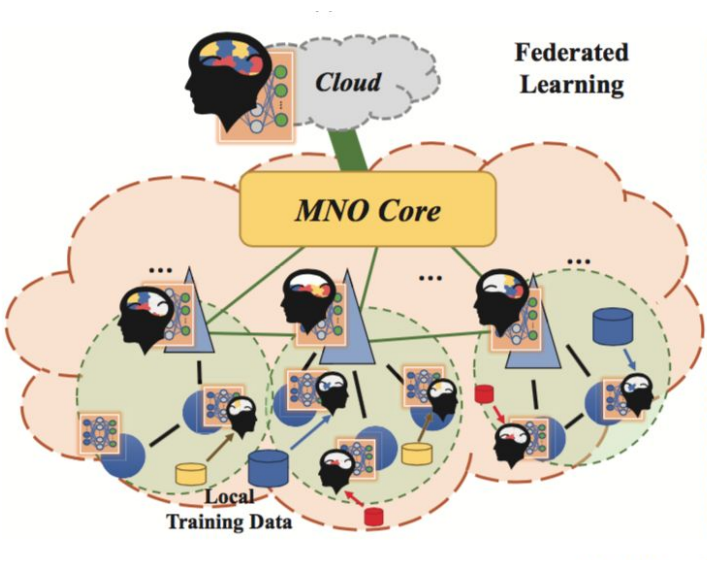
Case Study 1 - Traffic Engineering - The Solution



Centralized DRL



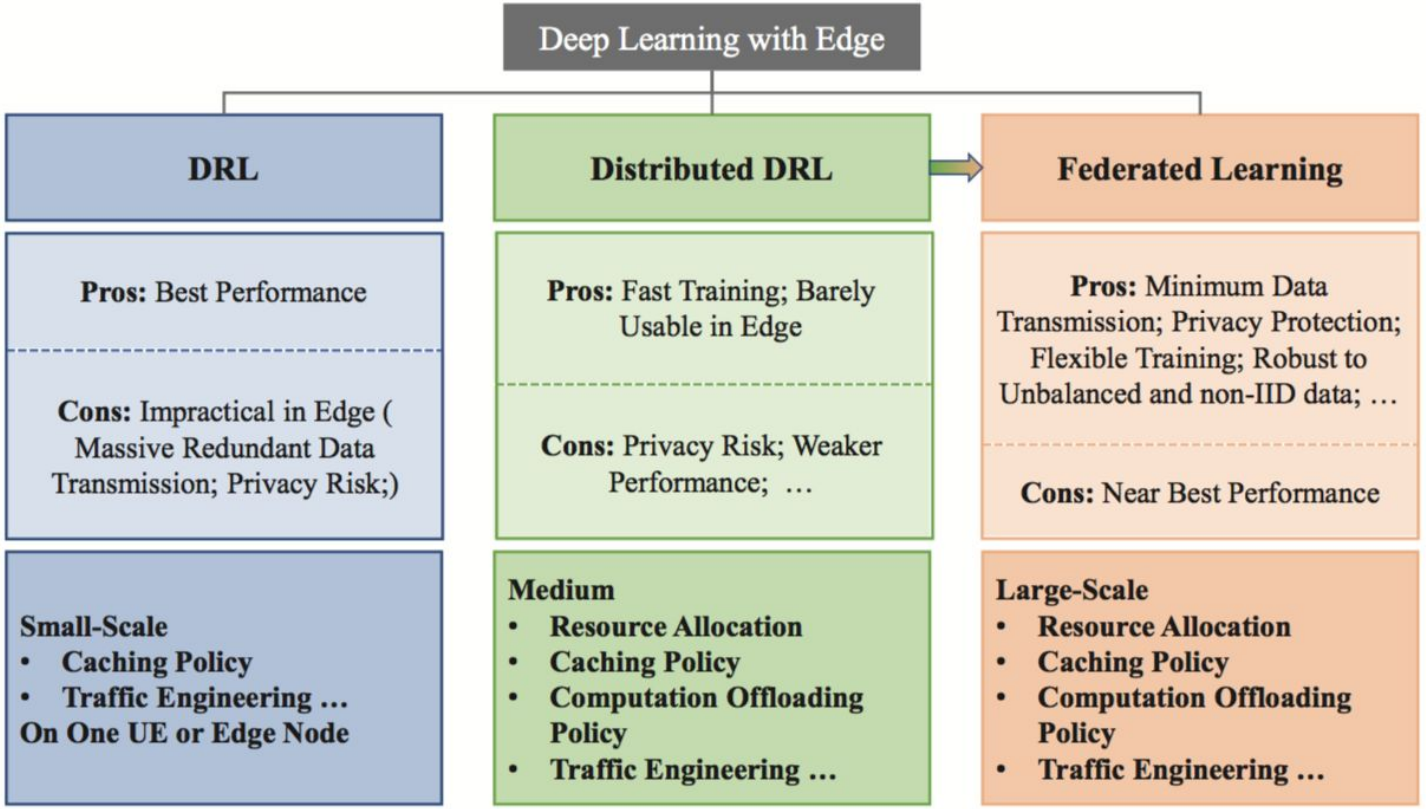
Distributed DRL



Distributed DRL + FL

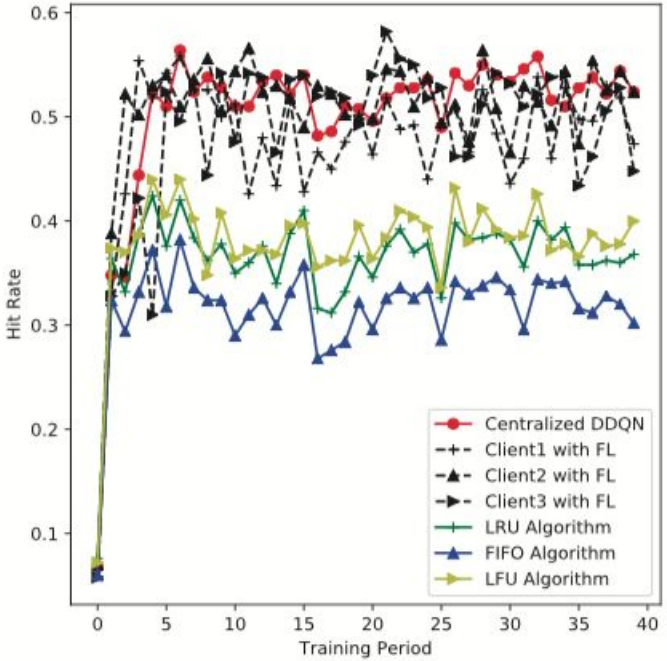
Source: https://arxiv.org/abs/1809.07857v1

Case Study 1 - Traffic Engineering - The Solution

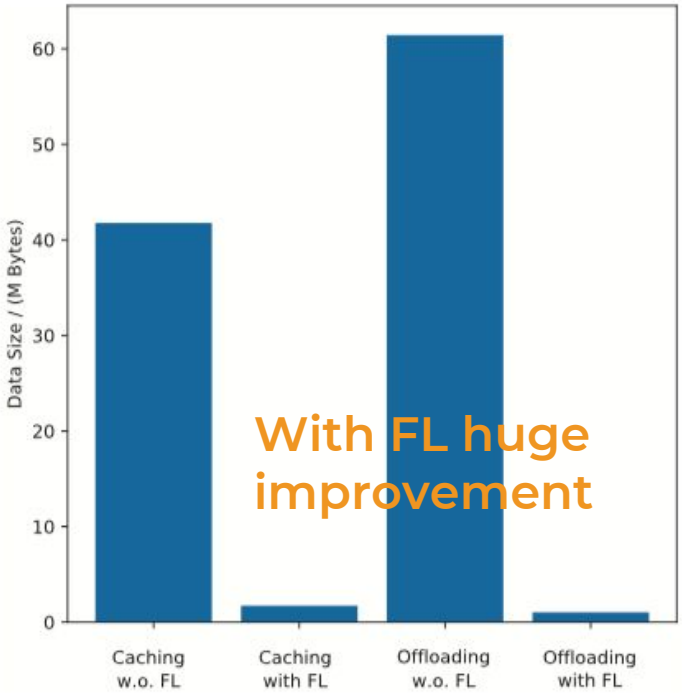


Source: https://arxiv.org/abs/1809.07857v1

Case Study 1 - Traffic Engineering - The Result



(a) Performance comparison in edge caching

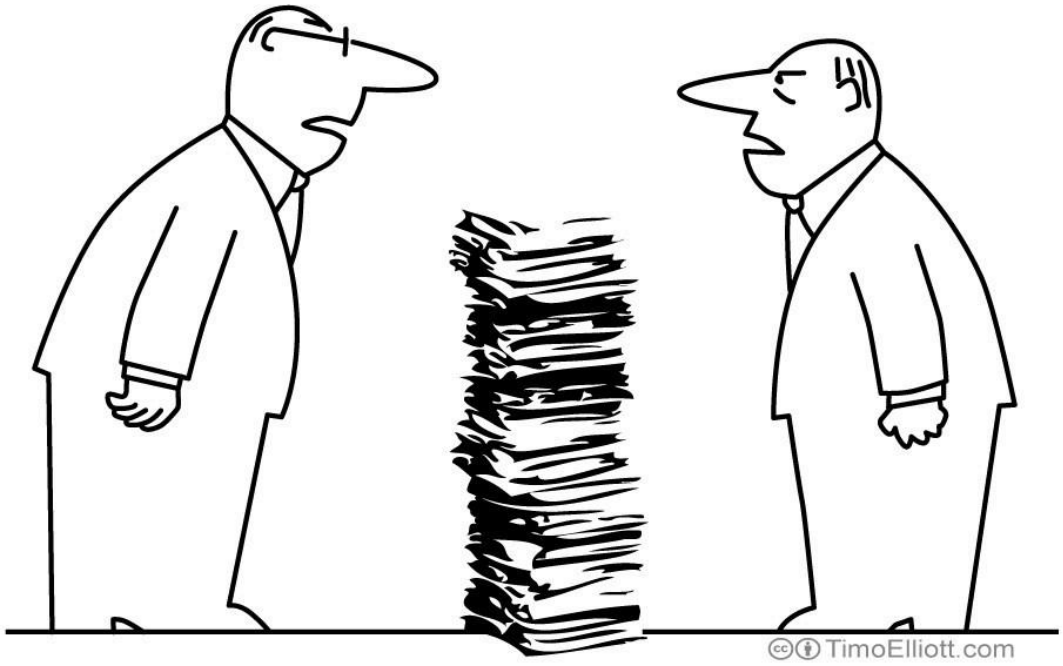


(d) Comparison of transmission cost

Source: https://arxiv.org/abs/1809.07857v1

Case Study 2 - Personal Data Anonymization - The Problem

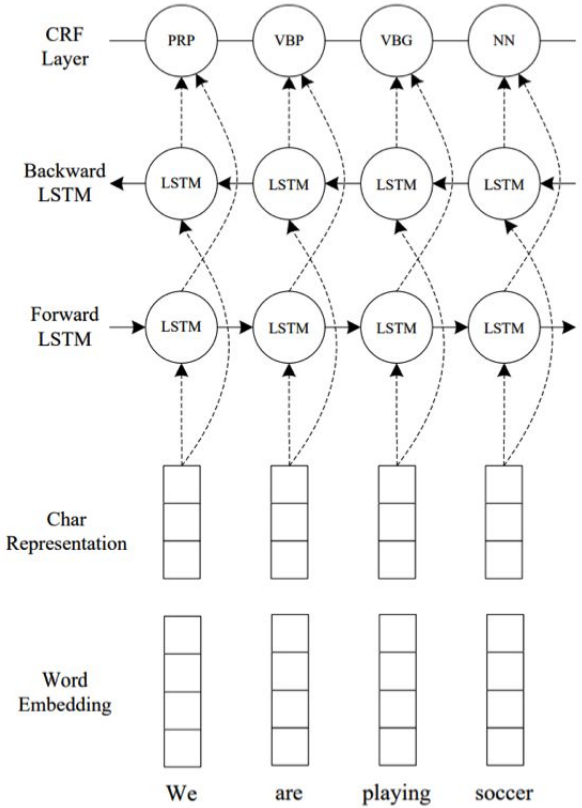
Personal Data



"No, it's MY data!"

Needs to be Anonymized at Edge before sending to Cloud

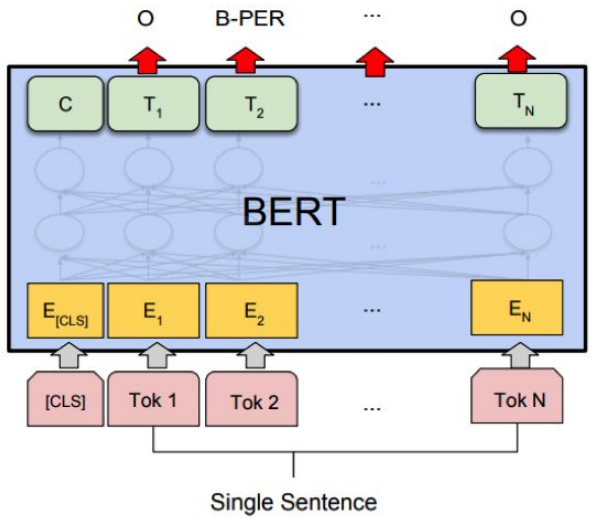
Case Study 2 - Personal Data Anonymization - The Solution



- **Named Entity Recognition(NER) using BLSTM+CRF**
- **Word embedding and char embedding**
- This model obtain a F1 score of 91.21 on CoNLL-2003 dataset.

Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): 2016. p. 1064–74.

Case Study 2 - Personal Data Anonymization - The Solution



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

- Named Entity Recognition with BERT.
- Feed the final hidden representation for each token into a classification layer over the NER label set.
- The predictions are not conditioned on the surrounding predictions (i.e., non-autoregressive and no CRF).
- This models shows a F1 score of 92.8 ON CoNLL-2003 dataset.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Case Study 2 - Personal Data Anonymization - The Results

Model	F1
Chieu and Ng (2002)	88.31
Florian et al. (2003)	88.76
Ando and Zhang (2005)	89.31
Collobert et al. (2011)‡	89.59
Huang et al. (2015)‡	90.10
Chiu and Nichols (2015)‡	90.77
Ratinov and Roth (2009)	90.80
Lin and Wu (2009)	90.90
Passos et al. (2014)	90.90
Lample et al. (2016)‡	90.94
Luo et al. (2015)	91.20
This paper	**91.21**

Table 5: NER F1 score of our model on test data set from CoNLL-2003. For the purpose of comparison, we also list F1 scores of previous top-performance systems. ‡ marks the neural models.

System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition results. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

Case Study 2 - Personal Data Anonymization - The Results

Data: Text and logs with 16 types of personal data

- Name, Email, address
- IP address, MAC address, host/server name
- Range 10 to 30 pages

No. of Documents	NER Time GPU (1 GPU)	NER Time CPU (4 core)	Factor
1	20 - 30 sec	60 - 120 sec	3 - 4X
100	30 - 45 mins	4 - 8 hrs	8 - 10X
1000	4 - 8 hrs	3 - 9 days	18 - 27X

Edge Computing - Opportunities /Challenges

1. Accelerating AI@dge Tasks by Edge Computing Systems
2. Efficiency of AI@dge for Real-time Mobile Communication System
3. Tight Federation among Mobile operators and service providers
4. Distributed Deep Learning and Deep RL frameworks to be evolved
5. AI@dge leveraging Transfer Learning, Adaptive Learning…

Re-inventing Edge Computing Apps Summary

$$\text{Edge Computing Apps} = f(AI,\ 5G, C)$$

$$\text{Where } C \text{ (compute)} = \{\ \text{GPU}, \text{CPU}, \text{Quantum}, \text{FPGA}\ \}$$

Thanks to NVIDIA & Manish Harsh

Q & A

Will Edge Eat the Cloud?

Gyana Dash
gyana.dash@gmail.com

Session Description

Significant breakthrough in 5G has evolved many IoT applications in various fields including business, manufacturing, health care and transportation. The evolution of GPU is the key enabler to the enriched applications by leveraging the power of AI @ the Edge. Edge computing still leverages the cloud as a crucial part of the ecosystem and many applications will harness the power of 5G features such as high speeds multi-gigabit connections, huge amounts of data bandwidth, unprecedented amounts of capacity, super-low latency and ultra-reliable low latency communications (URLLC). This session will explore the opportunities of some of the interesting applications to help our community and environment.

Abstract

As NVIDIA pioneers in proving Moore's law, the GPU enabled devices at the edge will have enough processing capability and power efficiency to run AI algorithms. Combined with the 5G evolution in the traditional mobile communication system and rapid AI innovations the edge computing applications will emerge to solve many interesting problems in various fields. Lightweight AI engines can be used at the edge for training and reasoning which is suitable for low-latency IoT services and can cover all ubiquitous intelligent edge applications.
The application of AI @ edge is still in the early stage and the coming years will be a critical period to harness the power of 5G and GPU for innovations that transforms our lives.

There are challenges to be solved both in 5G and AI, but potential solutions to the problem will lead to revolution in Edge Computing.
The Edge Computing ecosystem calls for security requirements and many organizations like ETSI MEC and OpenFog are working on security requirements and it will continue to evolve to address privacy, integrity and trust. In addition to security, location specific governance, regulations and compliance will emerge along with the evolution of Edge Computing frameworks and applications.