Advanced Topics in Service-Oriented Computing and Cloud Computing, Winter 2018
http://dsg.tuwien.ac.at/teaching/courses/socloud/

Service-oriented and Cloud Computing – Recap, Trends and Focus Points

Hong-Linh Truong
Faculty of Informatics, TU Wien

hong-linh.truong@tuwien.ac.at
http://www.Infosys.tuwien.ac.at/staff/truong
@linhsolar

Goals

- Understand current advanced developments and trends
  - With IoT, fog/edge computing, bigdata/data science, network functions and clouds
- Capture key questions for this course in the following 4 sub-areas
  - Scalable data, services and systems management
  - Elasticity and control
  - Big data analytics
  - End-to-end engineering analytics

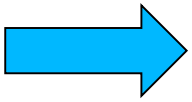
Identifying key focuses in this course

NEW TRENDS

Co-located Services in an ecosystem

- Big data, IoT Events/datahub, IoT device management, Serverless function, AI as a service, etc. are collocated in the same place
  - Google, Amazon, Azure, etc

- Multiple types of services provided by different providers in the same cloud infrastructures



Poly* aspect is a strategy:
https://www.thoughtworks.com/radar/techniques/polycloud

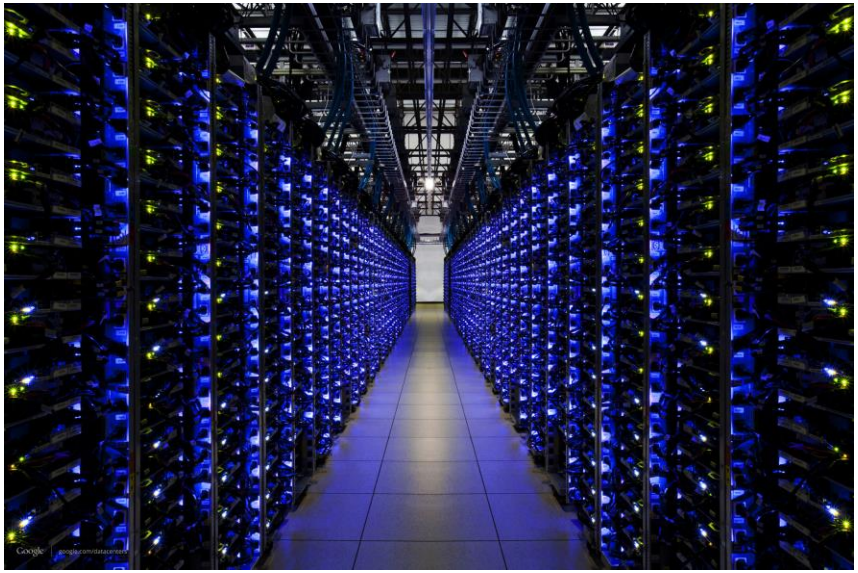
Multiclouds

- Multi aspects
  - Distributed, small and big, diverse types of resources and services
- Mini clouds and micro data centers
  - Used in edge/fog computing model to deploy resources close to data and analytics
- Cloud technologies: no longer associated with “big data centers”
  - They refer to “cloud” methodologies, models and techniques for distributed computing and services

Take a read: https://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2017-state-cloud-survey

Connecting data centers to IoT

Data Center: Processing, Storage, Networking, Management, Distribution



http://www.infoescola.com/wp-content/uploads/2013/01/datacenter-google.jpg

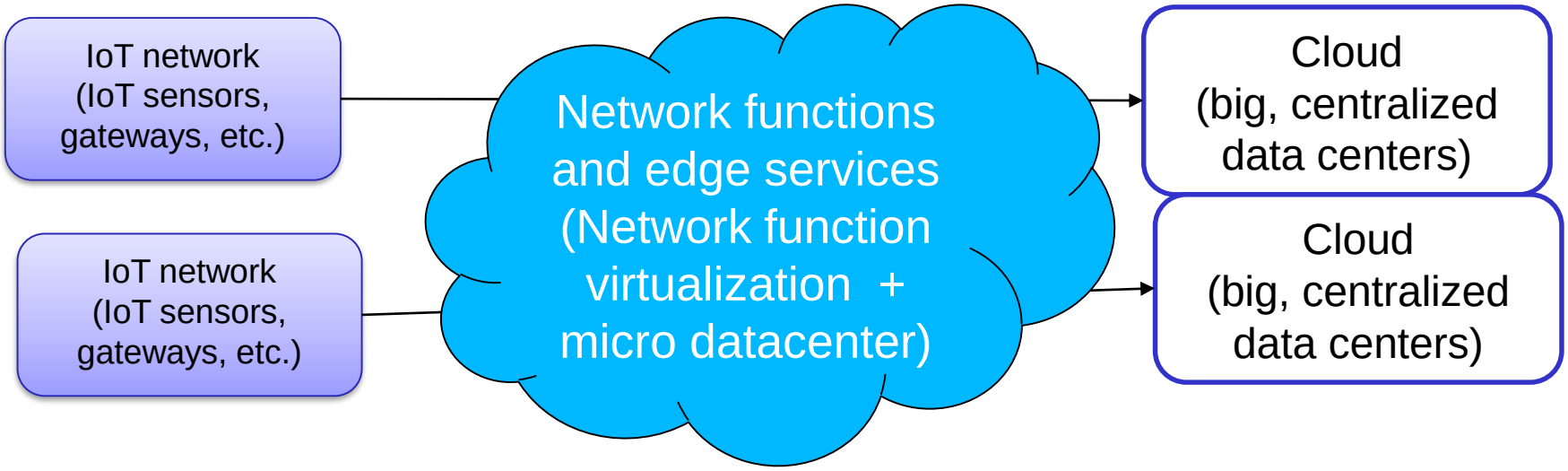
IoT devices: Gateways, Sensors, Actuators, Topologies of Gateways

http://www.control4.com/blog/2014/03/the-internet-of-things-and-the-connected-home

But there are many forms

Combining IoT + Network functions + Clouds

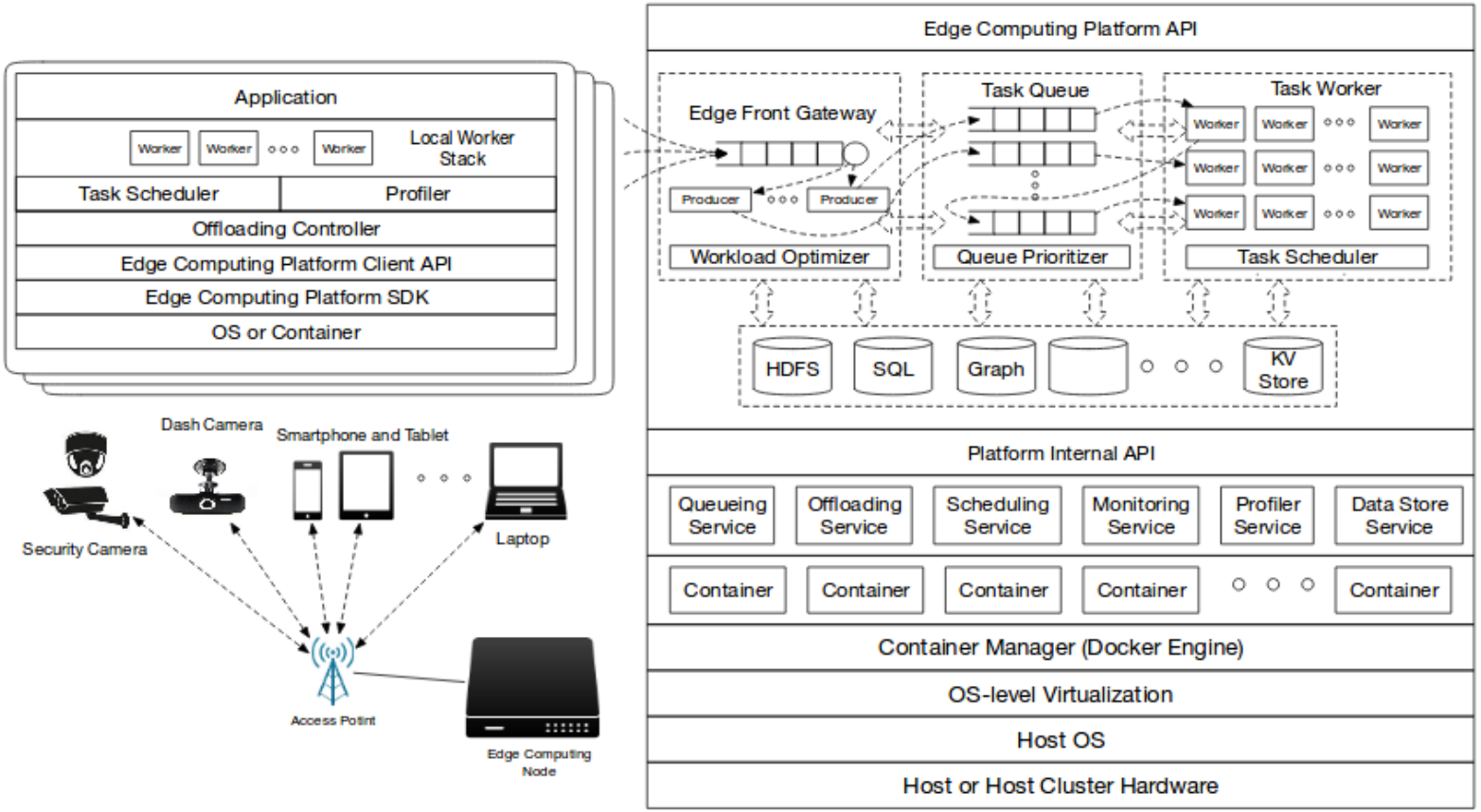
- Resources as services: IoT networks (sensors, gateways), edge/fog systems (micro data centers and network functions), big data centers (cloud VM, storage)



Just Googling edge/fog computing keywords to see how many papers have been published!
Note: look at papers from both computer science and communications

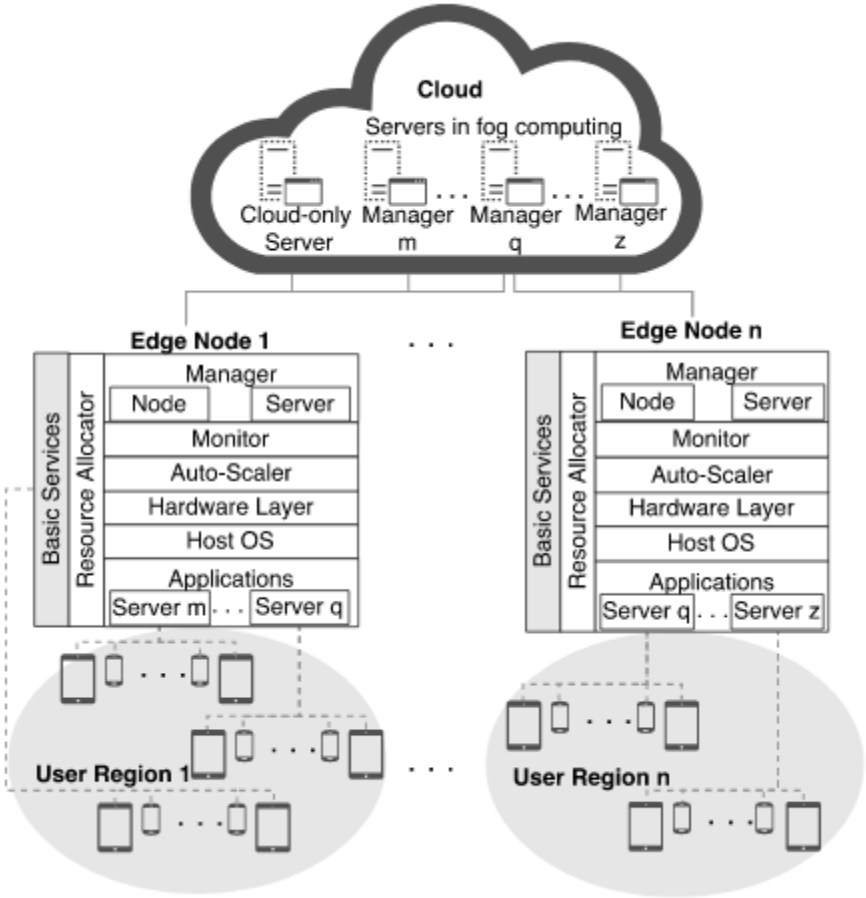
Some samples from fog/edge/cloud papers

From: Shanhe Yi, Zijiang Hao, Qingyang Zhang, Quan Zhang, Weisong Shi, and Qun Li. 2017. **LAVEA: latency-aware video analytics on edge computing platform**. In Proceedings of the Second ACM/IEEE Symposium on Edge Computing (SEC '17). ACM, New York, NY, USA, Article 15, 13 pages. DOI: https://doi.org/10.1145/3132211.3134459



Some samples of fog/edge/cloud papers

From: N. Wang, B. Varghese, M. Matthaiou and D. S. Nikolopoulos, **"ENORM: A Framework For Edge NOde Resource Management,"** in IEEE Transactions on Services Computing.
doi: 10.1109/TSC.2017.2753775



Services and Cloud for AI/IA & VR/AR

- Artificial Intelligence and Intelligence Amplification provided as services
- Connected *
  - Car, Health, and Assets
- Cloud robotics
- Manufacturing/Industry 4.0
- Virtual reality (VR) and augmented reality (AR) utilizing IoT data and services

Take a read: https://conficio.design/mixed-reality-and-iot

Technical trends

- Service interfaces
  - Not just REST, but also well-defined interfaces using messages via brokers and gRPC
- Service implementation
  - Java, Python, Nodejs, Go, etc.
- Underlying distributed resources
  - Raspberry PI, Virtual machines, OS Containers, virtual clusters, CPU cluster + GPU, Cloud TPUs
- “Microservices” mindset
  - Container, Container and Container! Kubernetes (K8s) for orchestration
  - Serverless Computing: Moving from pay-per-use “infrastructures” to pay-per-use function calls

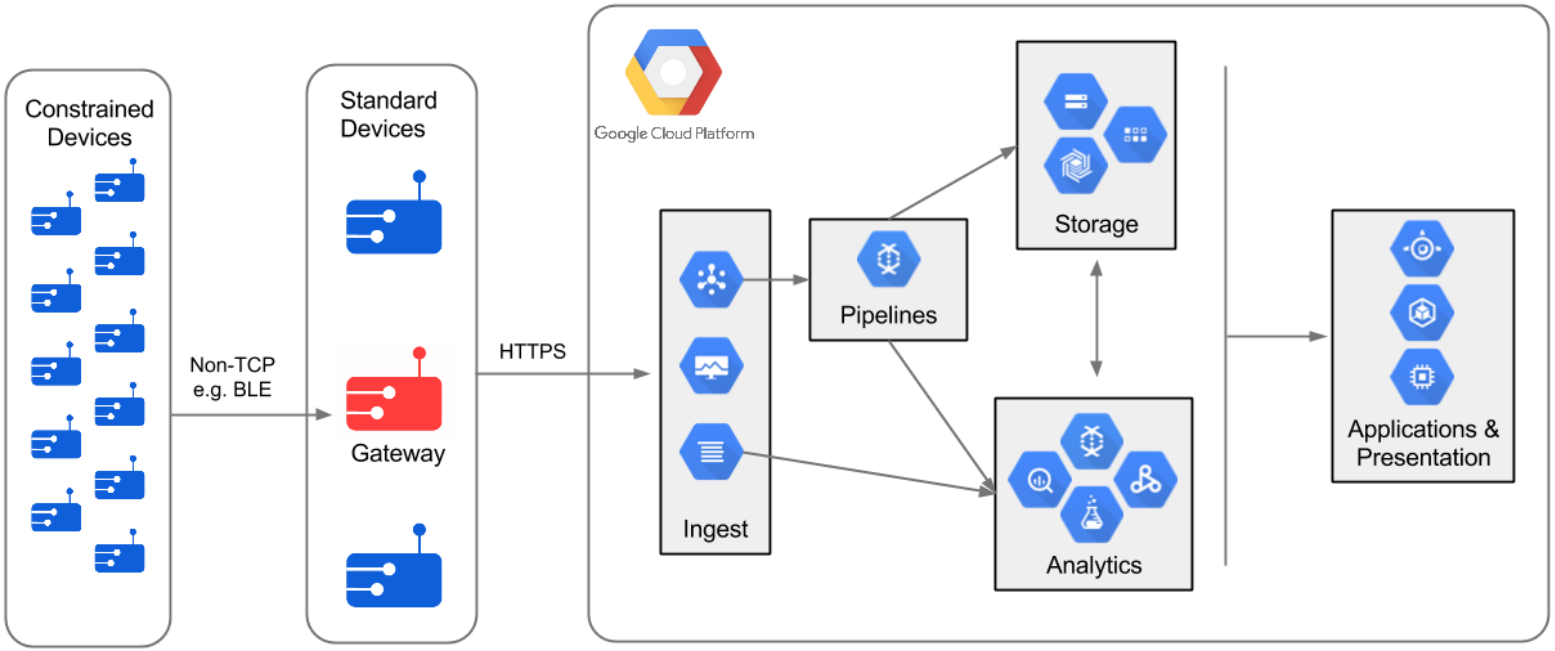
#1 key focus in this course

SCALABLE DATA, SERVICES AND SYSTEMS MANAGEMENT

Main types of virtualization of infrastructures for distributed apps

- Compute resource virtualization
  - Compute resources: CPU, memory, I/O, etc.
  - To provide virtual resources for „virtual machines“
- Storage virtualization
  - Resources: storage devices, harddisk, etc.
  - To optimize the usage and management of data storage
- Network Function Virtualization
  - Network resources: network equipment
  - To consolidate network equipment and dynamically provision and manage network functions

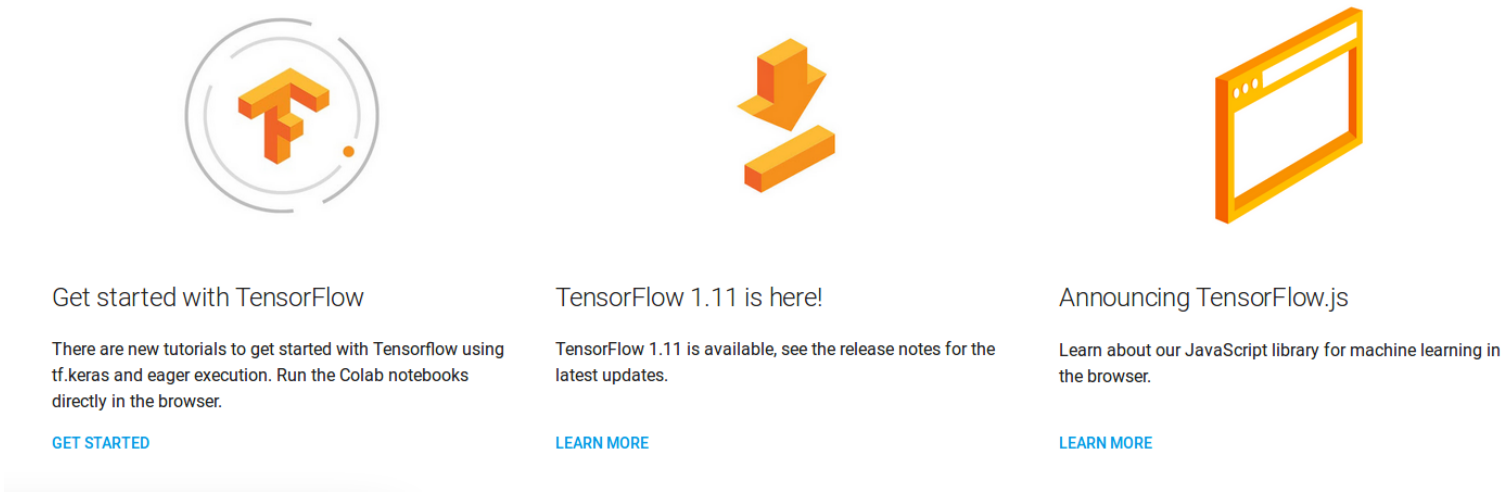
Example: IoT scenario



Source: https://cloud.google.com/solutions/architecture/streamprocessing

Example: AI with Tensorflow

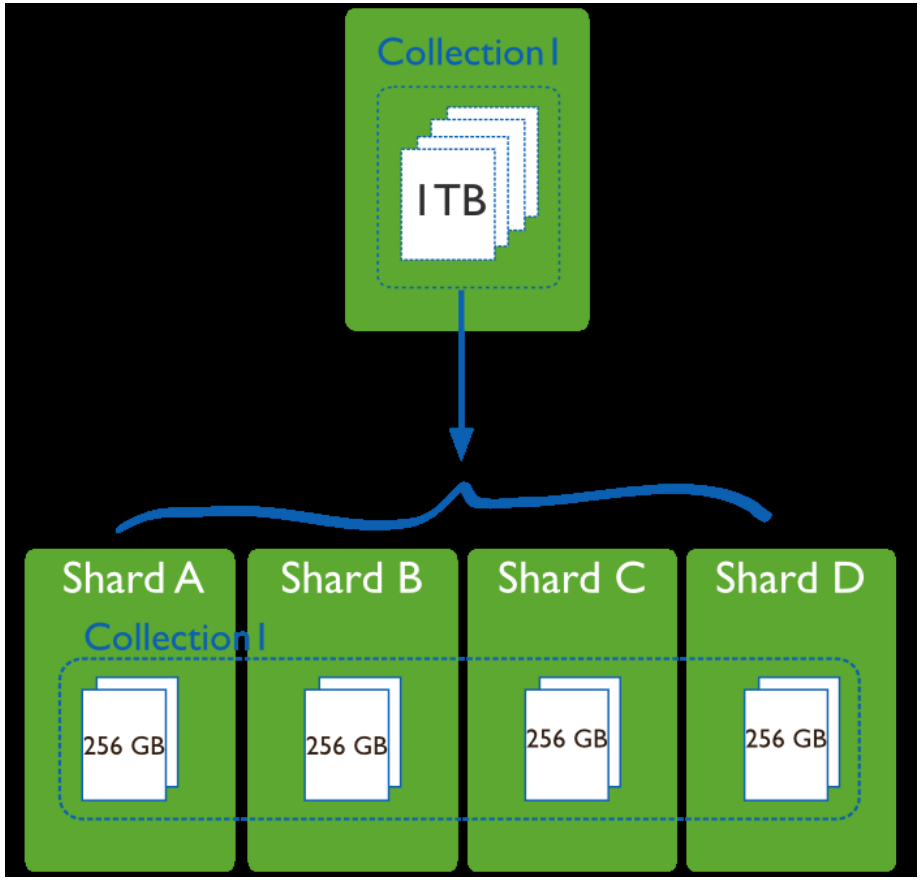
Key issue: algorithms (require a lot of computing power) + data (a lot of data)



Source: https://www.tensorflow.org/

Data Sharding

Need also
Routing, Metadata
Service, etc.



Source: https://docs.mongodb.org/manual/core/sharding-introduction/

Load balancing of data services

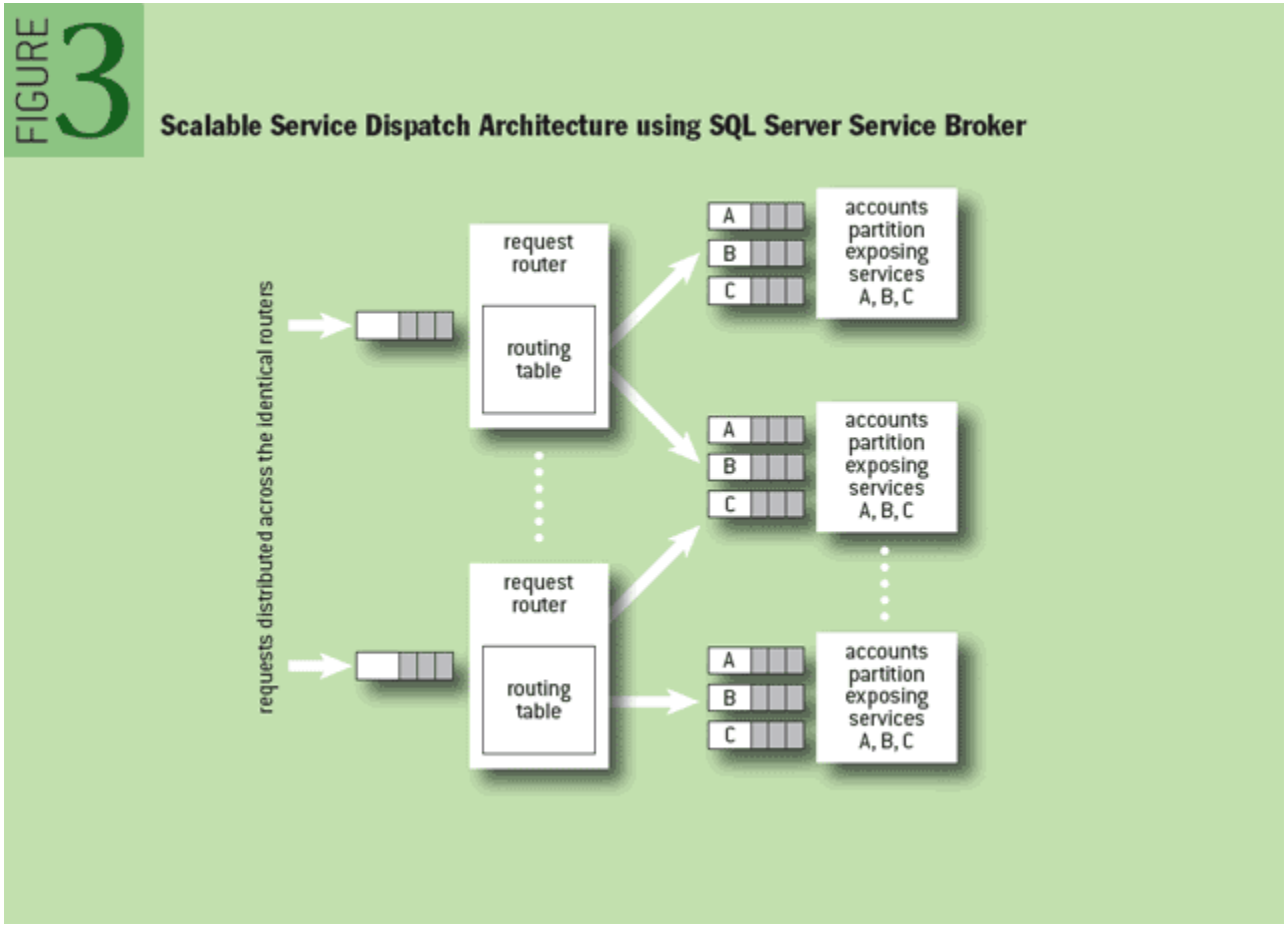
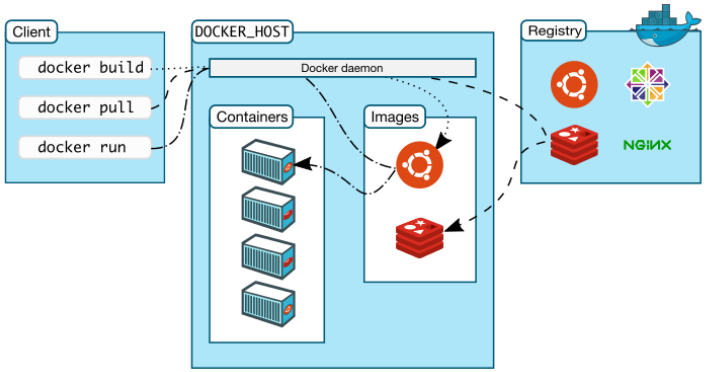
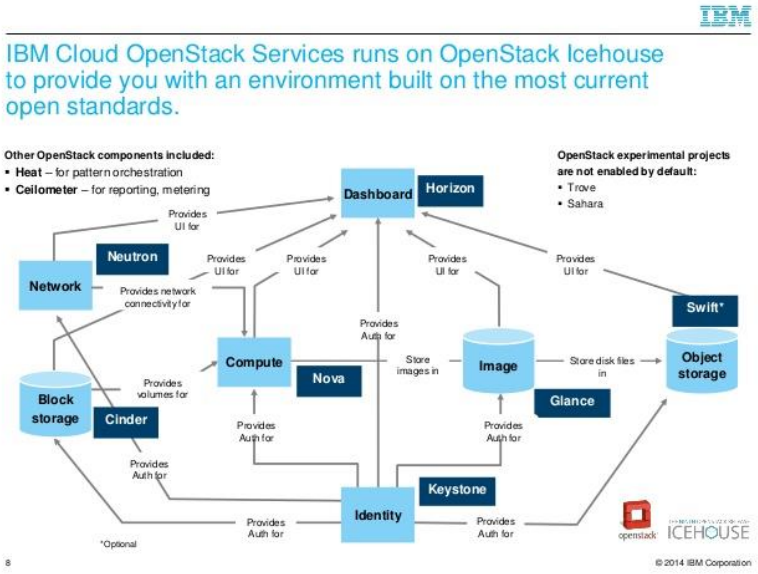


Figure source: http://queue.acm.org/detail.cfm?id=1971597

Examples



Source: https://docs.docker.com/engine/understand



Source: http://www.slideshare.net/OpenStack_Online/ibm-cloud-open-stack-services

Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.

It groups containers that make up an application into logical units for easy management and discovery. Kubernetes builds upon 15 years of experience of running production workloads at Google, combined with best-of-breed ideas and practices from the community.

Source: https://kubernetes.io/

Resource Management Across data centers/fog edge

How would we manage resources when services can be deployed across iot/edge and clouds?

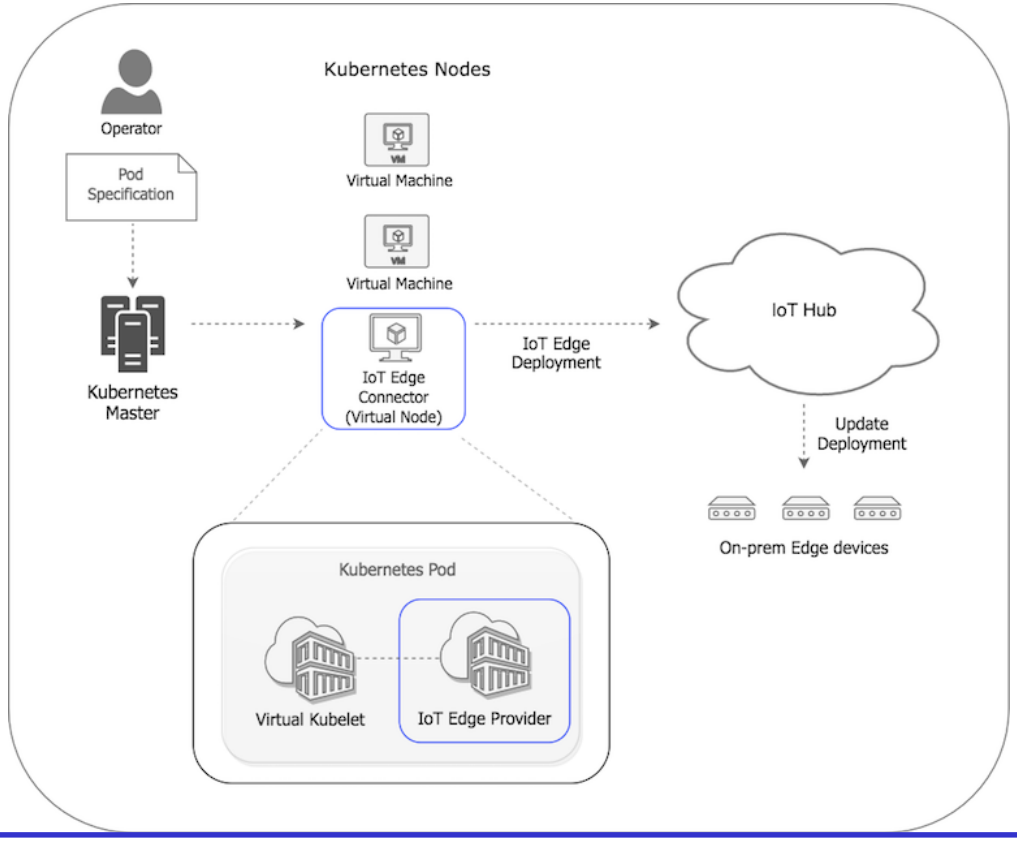


Figure source: https://github.com/Azure/iot-edge-virtual-kubelet-provider

Key research questions

- Management of complex types of resources (VM/containers, data, IoT devices, etc.)
  - Which are important advanced algorithms and design for scalable data/services/system management?
  - How to deal with high availability of data and data sharding?
  - How to deal with geographical multi-cloud load balancing?
  - Data services and container technologies

#2 key focus in this course

ELASTICITY AND CONTROL

Cloud Service Elasticity

A common problem in Cloud controls

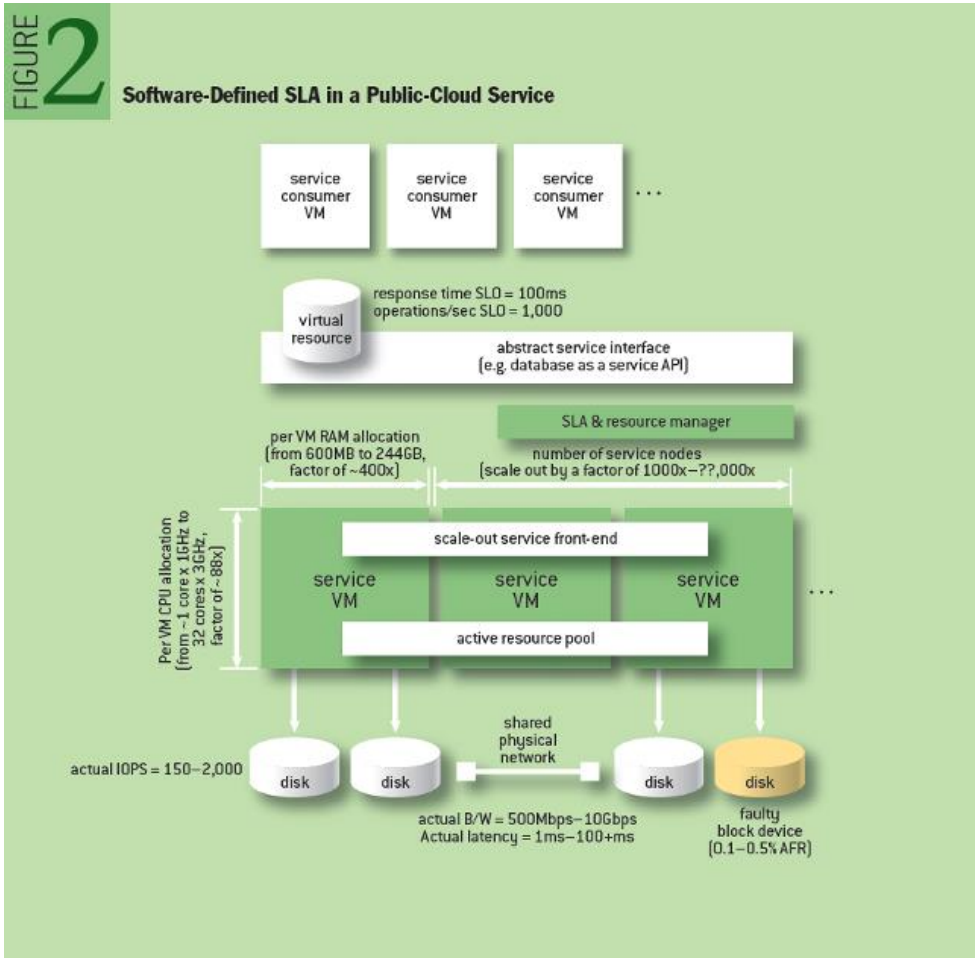


Figure source: http://queue.acm.org/detail.cfm?id=2560948

But we have much more complex situations

- Services in the car + services in the cloud
- Services in the robot + services in the cloud
- Interactions/communications
  - within cars/robots versus interactions/communications
  - among cars/robots and clouds
- Different data flows
  - Within the edge sites and among edge sties
  - Between the edge and the cloud

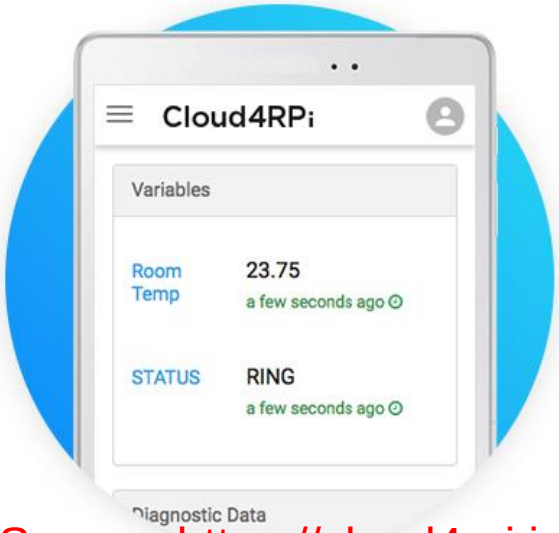
Edge/Fog Computing Execution Models

- “Off-loading styles”
  - Many papers taking about off-loading data from clouds or IoT data sources to different edge/fog resources for computing
    - Workflows, mobile off-loading computing, off-loading applications in edge computing
- “Reactive styles”
  - Data sources push data to different locations and data processing components react with the availability of data
  - Serverless/function-as-a-service, IoT data stream analytics, etc.

Controls between clouds, edge/fog and IoT

Source: https://en.wikipedia.org/wiki/File:Waymo_self-driving_car_front_view.gk.jpg

Source: https://spectrum.ieee.org/automaton/robotics/robotics-software/cloud-robotics

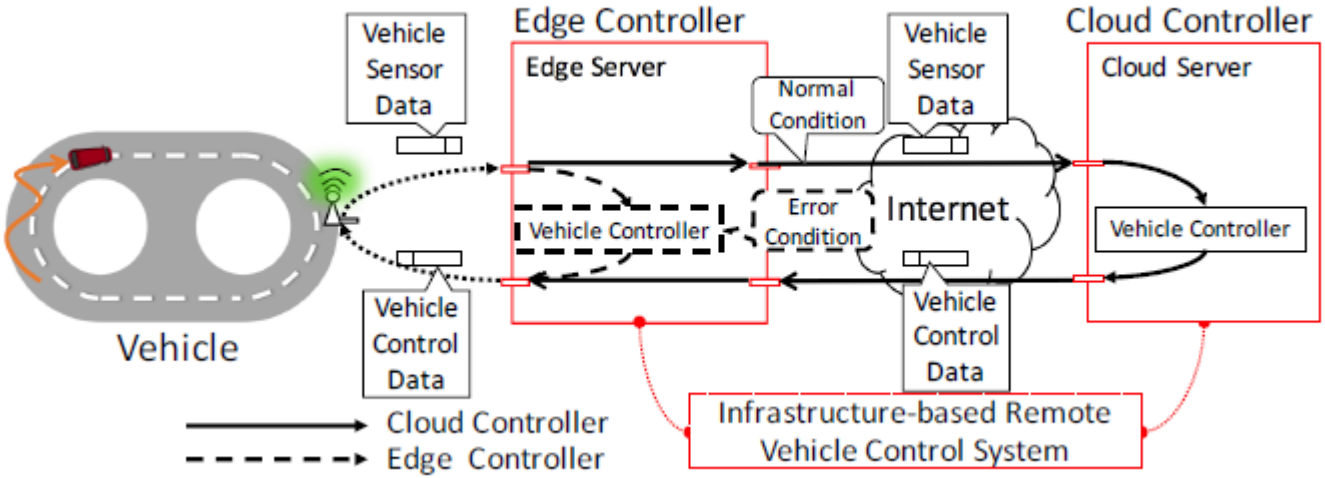


Source: https://cloud4rpi.io

How to control 1000+ assets or a network of Base Transceiver Stations?

Controls between clouds, edge/fog and IoT

Realtime control and analytics?



Source: K. Sasaki, N. Suzuki, S. Makido and A. Nakao, **"Vehicle control system coordinated between cloud and mobile edge computing,"** 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Tsukuba, 2016, pp. 1122-1127.
doi: 10.1109/SICE.2016.7749210

Key research questions

- Which algorithms and techniques can be used for elasticity controls in IoT, Cloud and fog computing?
- How to coordinate tasks among entities in IoT, edge systems and cloud centers?
- How does NFV leverage resources virtualization and elasticity techniques?
- Realtime/near realtime controls under uncertainty?

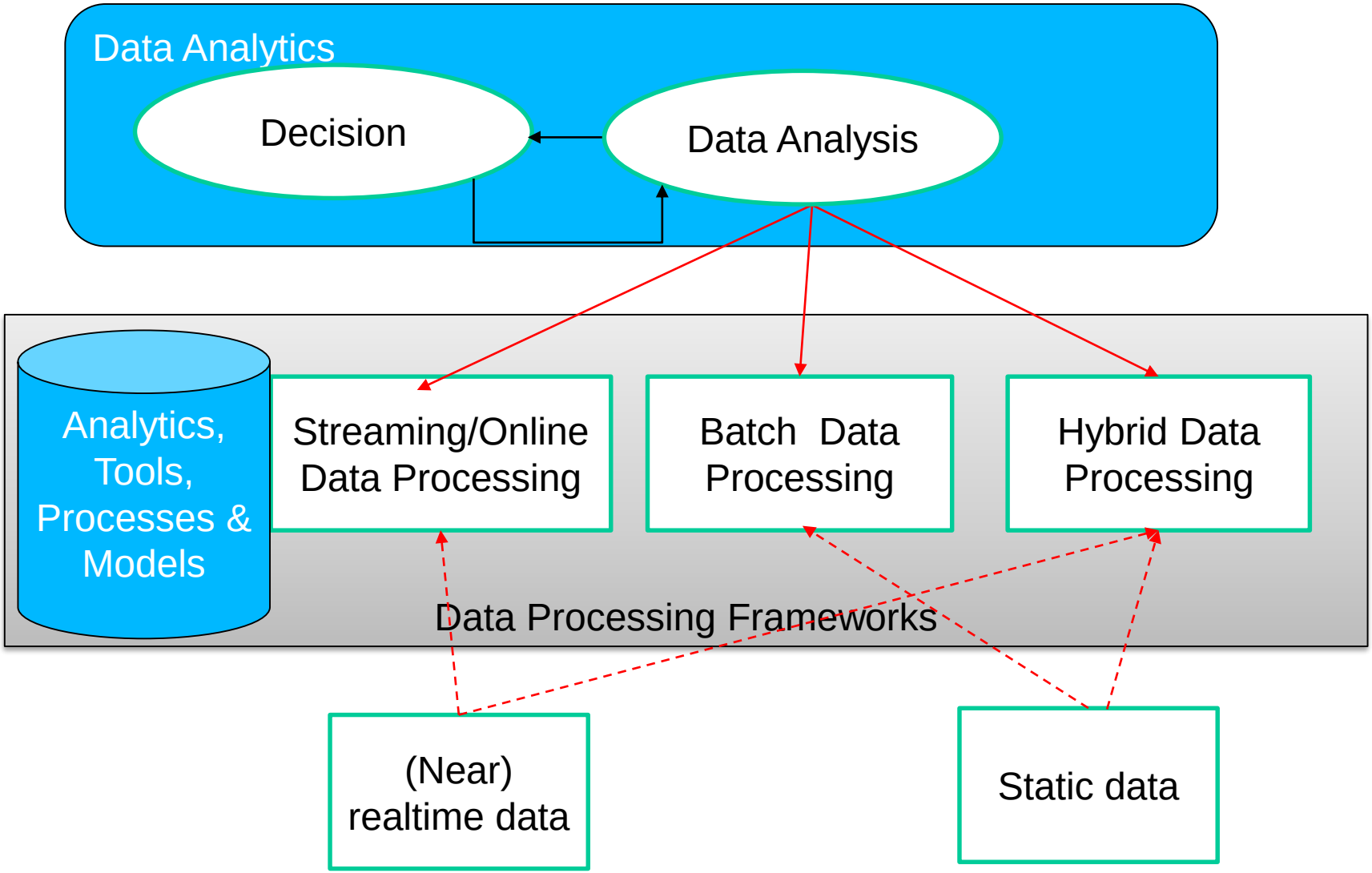
#3 key focus in this course

BIG SYSTEM FOR DATA ANALYTICS

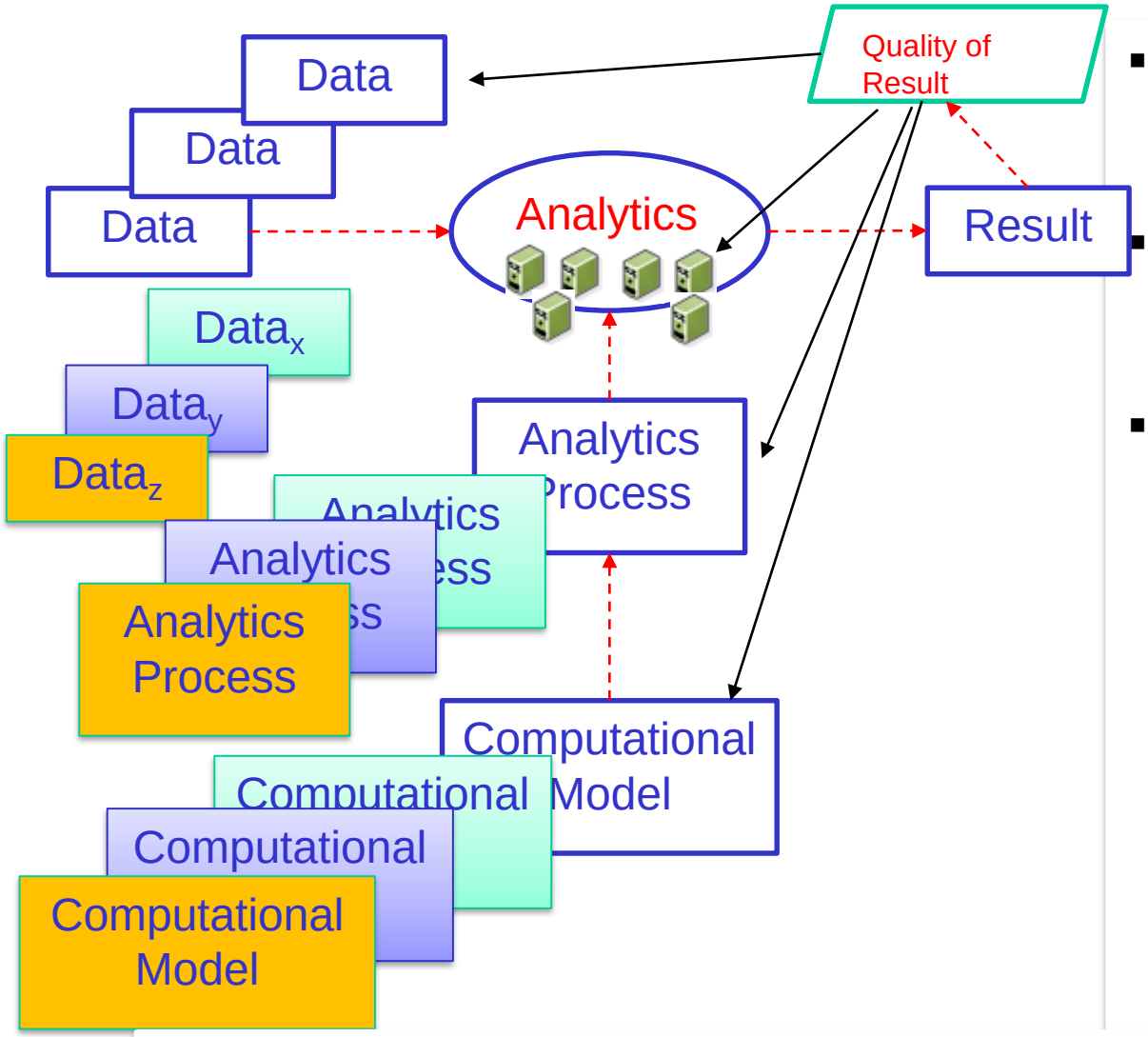
Data Processing Framework

- Batch processing
  - Mapreduce/Hadoop, Apache Spark
  - Scientific workflows
- (Near) realtime streaming data and complex event processing
  - Flint, Apex, Storm, etc.
- Hybrid data processing
  - Summingbird, Apache Kylin
  - Apache Spark

Conceptual View



Quality of analytics



- **More data** → more computational resources (e.g. more VMs)
- **More types of data** → more computational models → more analytics processes
- Change **quality of analytics**
  - Change quality of data
  - Change response time
  - Change cost
  - Change types of result (form of the data output, e.g. tree, visual, story, etc.)

Hong-Linh Truong, Schahram Dustdar, "Principles of Software-defined Elastic Systems for Big Data Analytics", (c) IEEE Computer Society, IEEE International Workshop on Software Defined Systems, 2014 IEEE International Conference on Cloud Engineering (IC2E 2014), Boston, Massachusetts, USA, 10-14 March 2014

Complex Machine learning flows

When GPU + Machine learning + Workflows are executed in complex systems for big data analytics

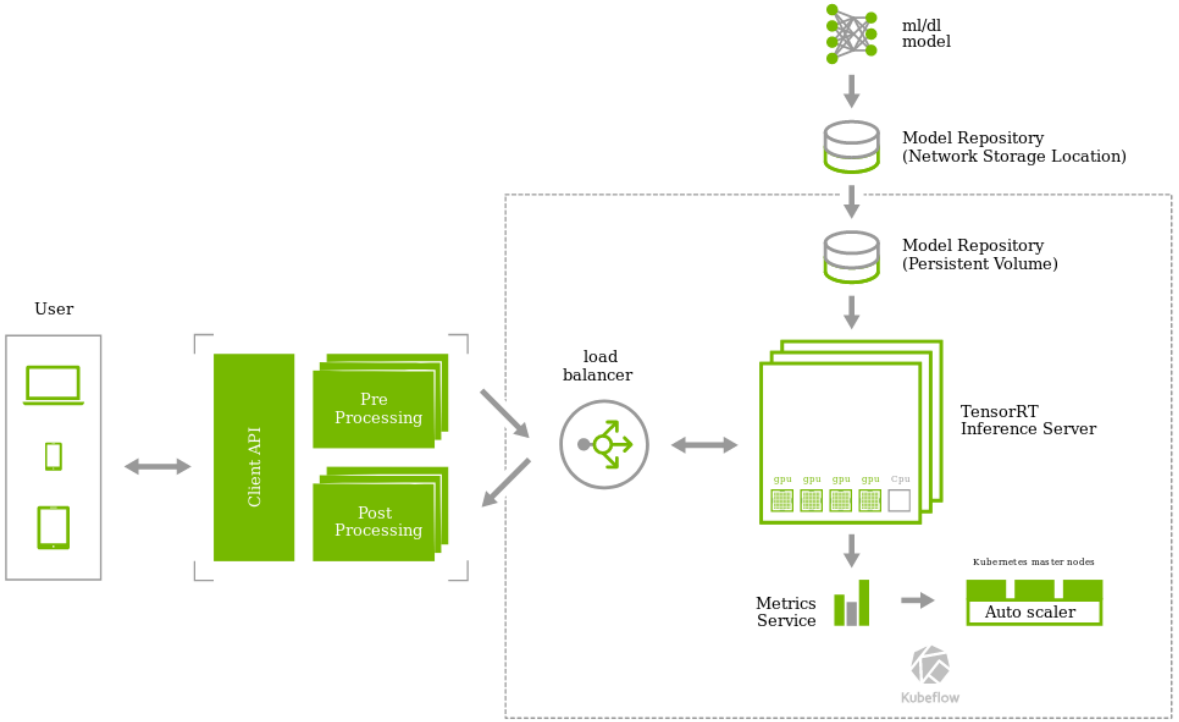


Figure source: https://www.kubeflow.org/blog/nvidia_tensorrt/

Human + Machine for Data Analytics?

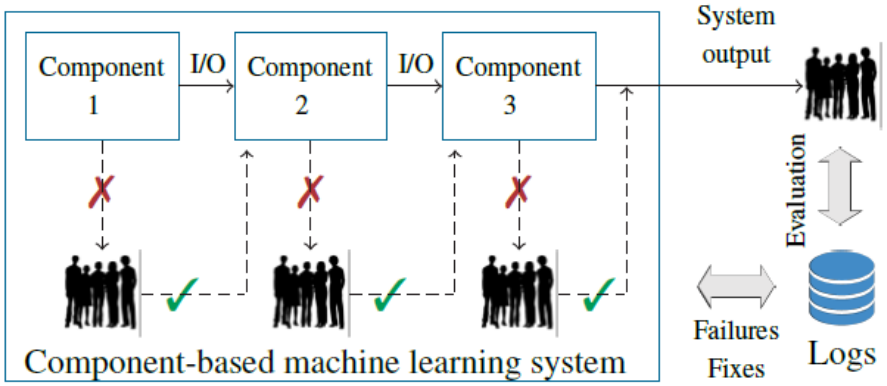


Figure 1: Troubleshooting with humans in the loop

Figure source: Besmira Nushi, Ece Kamar, Donald Kossmann and Eric Horvitz. On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems, AAAI 2017.

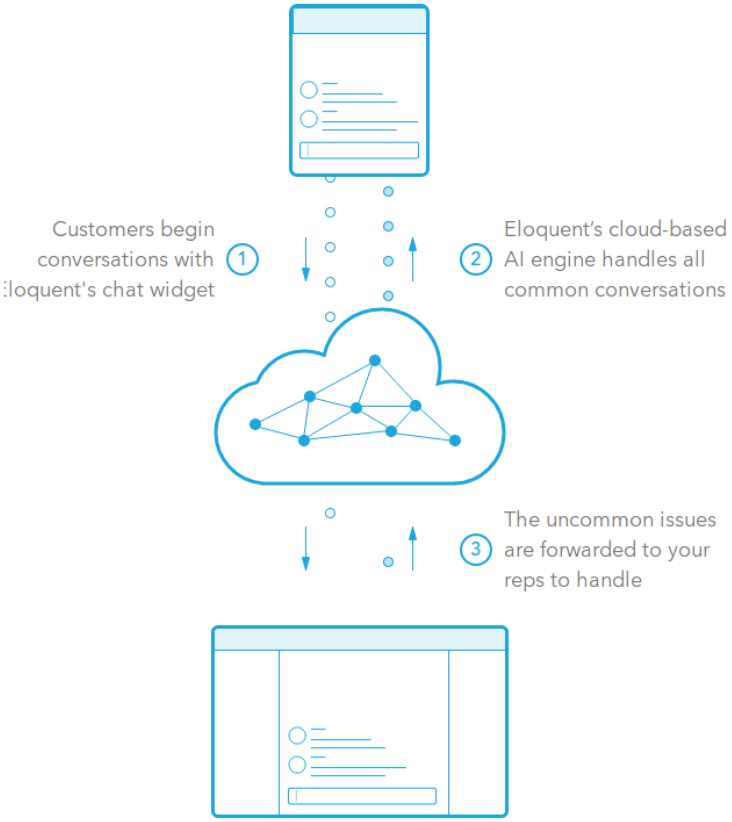


Figure source: https://www.eloquent.ai/

Key questions

- Which techniques and algorithms are important for big data ingest and analytics?
- How does data analytics leverage elasticity?
- How do big data analytics systems support machine learning?
- How to achieve quality-aware analytics?

#4 key focus in this course

Key focus in this course

END-TO-END SERVICE SYSTEM ENGINEERING

Common goals for IoT Cloud service engineering analytics

- Type 1
  - **Mainly focus** on IoT networks: sensors, IoT gateways, IoT-to-cloud connectivity (e.g., connect to predix.io, IBM Bluemix, Amazon IoT, etc.)
- Type 2
  - **Mainly focus** on (public/private) services in data centers: e.g., load balancer, NoSQL databases, and big data ingest systems
  - Using both open sources and cloud-provided services
- Type 3
  - **Equally focus** on both IoT and cloud sides and have the need to control at both sides
  - Highly interactions between the two sides, not just data flows from IoT to clouds

End-to-End resource management

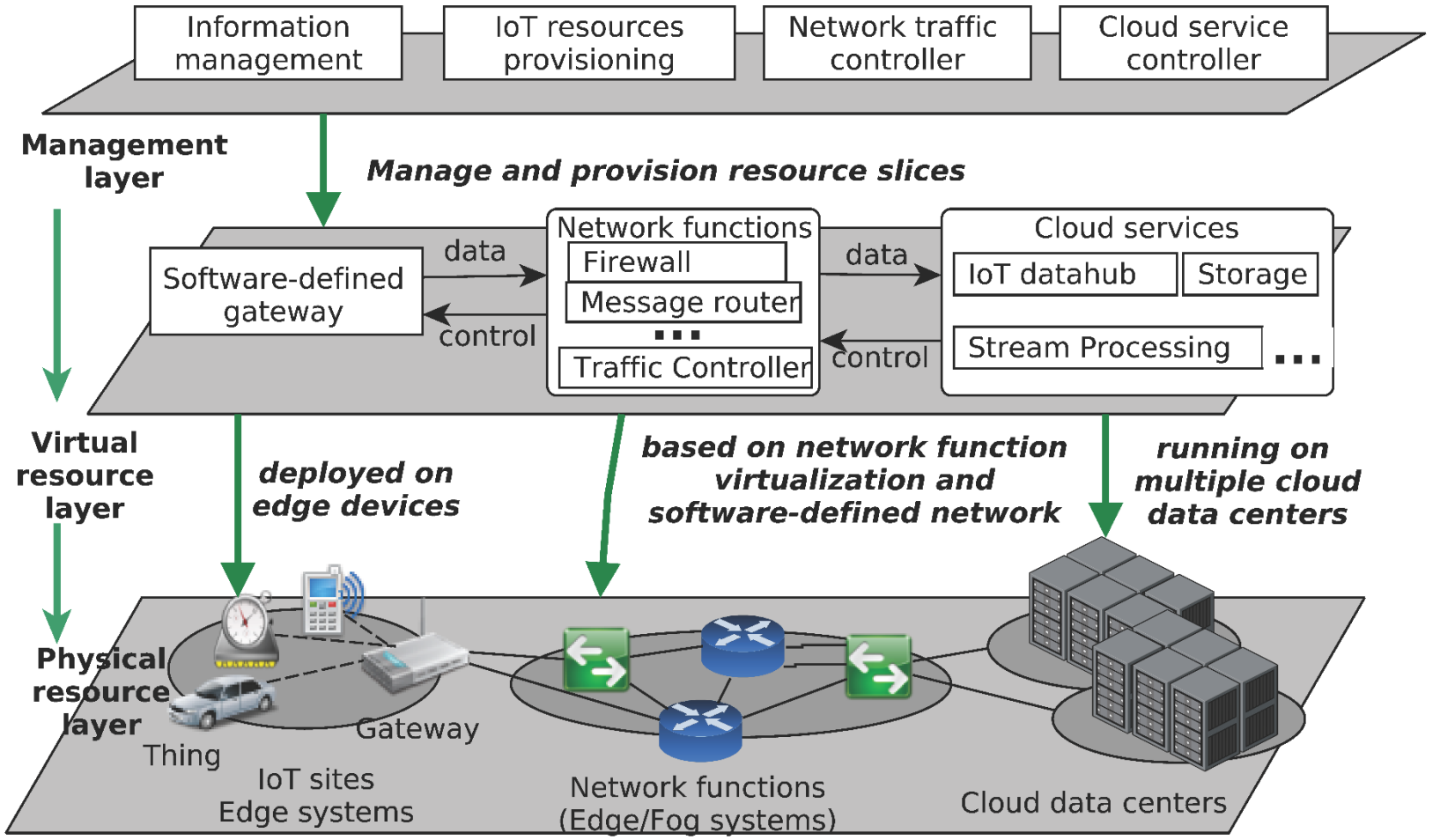
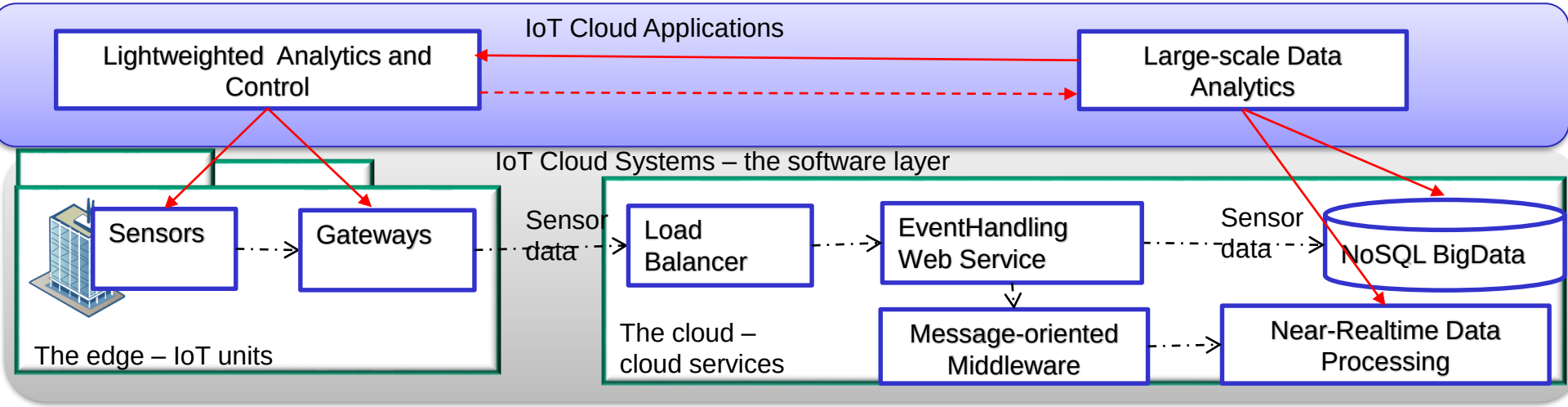
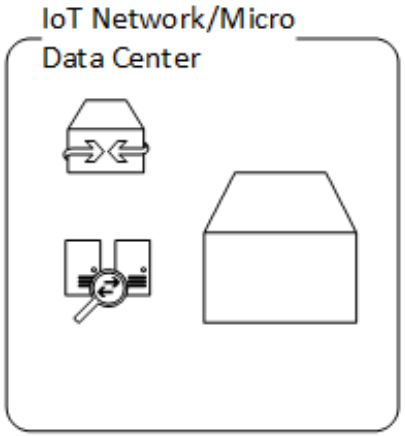
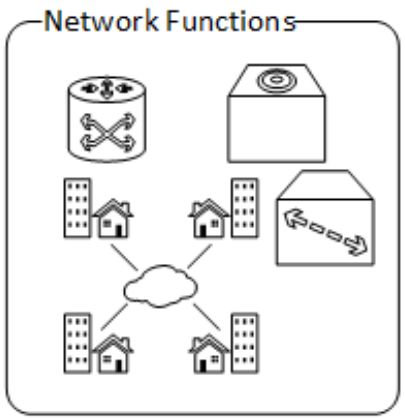


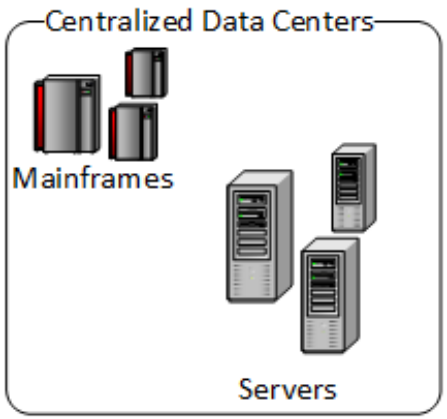
Figure source: Duc-Hung Le, Nanjangud C. Narendra, Hong Linh Truong:
HINC - Harmonizing Diverse Resource Information across IoT, Network Functions, and Clouds. FiCloud 2016: 317-324

End-to-End View on Applications and Systems









Uncertainty & testing

- Many possible uncertainties associated with interactions among data, services and systems

- Supporting testing uncertainties and uncertainties analytics
  - Conventional aspects, e.g., infrastructural resources and typical service/system operations
  - Emerging novel aspects: data uncertainties (data/data-centric), elasticity of resources (w.r.t function and composition), and governance (related to business/trustworthiness)

Key questions

- How do we monitor and analyze IoT, cloud and fog/edge systems?
- Which are important techniques and tools for instrumenting and monitoring IoT, cloud and fog systems?
- How do we determine metrics for end-to-end views?
- What are important types of uncertainties? How to measure and test them?

Summary

- Different types of services built atop IoT, fog/edge and cloud systems
  - The underlying infrastructures and systems are very complex
  - Virtualization and elasticity are important techniques
  - Big data problems: dealing with a lot of near real time data
  - Performance monitoring and testing
- The focus points
  - Advanced system design techniques and algorithms
  - In the next 3 weeks: read many papers about the current state-of-the-art of IoT/fog/edge cloud systems

Thanks for your attention

Hong-Linh Truong
Faculty of Informatics
TU Wien
truong@dsg.tuwien.ac.at
http://www.infosys.tuwien.ac.at/staff/truong