Edge Computing Research Challenges for IoT

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

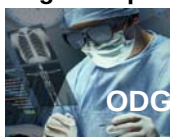
Today: Direct to Cloud

Drones

Static & Vehicular Sensor Arrays

Microsoft Hololens

Magic Leap



AR/VR Users

Tier 3

Internet

Tier 1 aka "Cloud"

Edge Computing for IoT

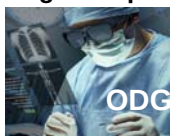


Drones

Static & Vehicular Sensor Arrays

Microsoft Hololens

Magic Leap



AR/VR Users

5G / 4G
Wi-Fi /
tethered

low-RTT
high-bw
network

Internet

Tier 3

Tier 2 aka “Cloudlets”

Tier 1

“Cloudlet” = Tier 2 Compute Infrastructure

aka “micro data center”, “edge cloud”, “fog node”

Small data center at the edge of the Internet

- **many sizes & form factors**
- **one wireless hop (5G / 4G / Wi-Fi / other + fiber / LAN) to tier 3**
- **multi-tenant, as in cloud**
- **good isolation and safety (VM-based guests or containerized apps)**

Non-constraints **(relative to Tier 3)**

- **energy**
- **weight/size/heat**
- **extreme low cost**

Won't Mobile Devices Get More Powerful?

Year	Typical Tier-1 Server Processor	Speed	Typical Tier-3 Device Device	Speed
1997	Pentium II	266 MHz	Palm Pilot	16 MHz
2002	Itanium	1 GHz	Blackberry 5810	133 MHz
2007	Intel Core 2	9.6 GHz (4 cores)	Apple iPhone	412 MHz
2011	Intel Xeon X5	32 GHz (2x6 cores)	Samsung Galaxy S2	2.4 GHz (2 cores)
2013	Intel Xeon E5-2697v2	64 GHz (2x12 cores)	Samsung Galaxy S4	6.4 GHz (4 cores)
			Google Glass	2.4 GHz (2 cores)
2016	Intel Xeon E5-2698v4	88.0 GHz (2x20 cores)	Samsung Galaxy S7	7.5 GHz (4 cores)
			HoloLens	4.16 GHz (4 cores)
2017	Intel Xeon Gold 6148	96.0 GHz (2x20 cores)	Pixel 2	9.4 GHz (4 cores)

Source: Adapted from Chen [3] and Flinn [8]
"Speed" metric = number of cores times per-core clock speed.

Fig. 2. **The Mobility Penalty: Impact of Tier-3 Constraints**

Research Challenges at Tier 2

Security and Privacy

Overcoming Limited Elasticity

Dynamic Data Aggregation and Reconfiguration

Security and Privacy

How do we

1. Withstand vulnerability of Tier 2 to physical attacks?
2. Balance value of real-time sensor data (e.g., video) with privacy?
3. Leverage geospatial proximity for trust?
4. Avoid security/privacy delays in latency-critical paths?

Overcoming Limited Elasticity

How do we

1. Allocate limited resources at Tier 2 (relative to Tier 1)?
2. Handle flash crowds and data-driven spikes in demand?
3. Learn user and application behavior, and remember them for the future?
4. Recruit help (other cloudlets) rapidly and seamlessly when needed?

Dynamic Data Aggregation and Reconfiguration

How do we

1. Perform spatial and temporal aggregation for scalability (megasensor)?
2. Do real-time machine learning for optimal video encoding?
3. Avoid Tier 2 processing (and hence offered load) whenever possible?
4. Switch levels of aggregation rapidly and easily (hence be context sensitive)?