Edge Computing: The What, the Why, and the How

Aruna Balasubramanian (Stony Brook University)
Michael Rabinovich (Case Western University)

Where do we draw the edge?

Multiple definitions for Edge

1. "Extreme-edge computing": mobile devices are part of the edge
2. Edge is the first point of attachment for devices. For example, for wearable devices, smartphone is the edge. For smartphones, a fog server is the edge
3. Edge is the first point of attachment to the wired network (base station)
4. Edge is any point that is "much closer" to the end-device than to the cloud (CDNs' edge servers)

Take-aways

- Whenever writing about edge computing, be extremely clear about your definition
- PDs and PC chairs need to raise awareness among reviewers of different preconceptions
- The community would benefit from developing common terminology

What are the system assumptions about the edge?

- There is space for both application-specific and general edge computing deployments
- For app-specific deployments, capacity assumptions depend on app requirements
- Example of general purpose edge deployments:
  - A set of apps using CCTV deployment: surveillance, emergency response, traffic control apps
  - AR/VR applications
- Design considerations
  - The density of deployment
  - Who is paying for the edge infrastructure? AWS? Akamai? AT&T? Stand-alone organizations?
  - A chicken-and-egg problem for edge-native apps:
    - Critical mass of deployment is needed to justify creation of edge applications.
    - Critical mass of apps needed to justify deployment.

Is research on efficient live migration impactful in the context of edge computing?

- If one uses edge computing to keep data private on infrastructure they control, you don't need migration
  - Home-based applications
- VR/AR applications tolerate 20 ms delay. A single metro-area datacenter will work with 5G - no need for migration
  - E.g., Netflix has data centers within a regional ISP
  - E.g., CDNs have data centers with direct connections to multiple ISPs
- Surveillance and video analytics applications.
  - A single metro-area datacenter may not suffice due to bandwidth consumption - but still no need for live migration
- Possible motivations for live migration
  - Load balancing within the metro-area datacenter (similar to cloud issues)

Take-away:
  - Could not come up with a compelling use case for live migration in edge computing

Is research on automatic partitioning of apps between mobile device and the edge impactful?

Desirable:
- The environment can be different, so an edge-native app would be too inflexible.
- The conditions may change so there is an advantage in giving the app the flexibility to adapt

But difficult:
- Compromises reasoning about failures.
- Complicates costs prediction - we don't know where things will run hence how much they will cost.
- Fixed egde-native application is more deterministic and easier to reason about, for both users (can/can't run vs. graceful degradation) and developers

Sounds like a good research problem!
- Adding monitoring/bookkeeping to help with diagnostics?
- Defining clean abstractions to formulate policies on partitioning decisions and help with reasoning?

In search for killer app

Is it vehicle-to-vehicle and vehicle-to-edge computing?
- Why is Tesla not deploying it?
- Can I trust what other cars are telling me?
- What is value-add beyond vehicle-to-cloud?

Counterpoint: The vehicle-to-vehicle can be used as a redundant path for reliability. The companies may not be doing this but as researchers we should.

Standard stand-bys:
  AR/VR, 360 Video, video surveillance/analytics

A new entrant: Spectrum sensing and allocation.

- Needs millisecond latency.
- How often a given spectrum is occupied?
- How many of the devices using the spectrum are mobile?