# Edge Computing: the Journey Ahead

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

Living Edge Lab

Carnegie Mellon University

# ~$16 Billion Market in 2023

**"The global edge computing market size was valued at $15.96 billion in 2023 and is projected to grow from $21.41 billion in 2024 to $216.76 billion by 2032..."**

**Source: Fortune Business Insights**


**"The global edge computing market size was valued at USD 16.45 billion in 2023 and is expected to grow at a compound annual growth rate (CAGR) of 36.9% from 2024 to 2030. "**

**Source: Grand View Research**

***What is driving this order of magnitude growth?***

***What challenges does edge computing face?***

***Where will the big wins come from?***

***How will edge computing transform life worldwide?***

# What is Driving this Growth?

| | |
|---|---|
| **1. Cost-effective IoT data analytics**<br>*"Scalable real-time sensor analytics"* | **Bandwidth**<br>(peak and average) |
| **2. Highly responsive cloud services**<br>*"New applications and microservices"* | **Latency**<br>(mean and tail) |
| **3. Privacy-sensitive IoT Deployments**<br>*"Crossing the IoT Chasm"* | **Privacy**<br>(control of sensor data) |
| **4. Internet disruptions**<br>*"Disconnected operation for cloud services"* | **Availability**<br>(UPS for cloud) |
| **5. Data sovereignty**<br>*"In-domain storage and processing"* | **Legal/Business**<br>(bring compute to the data) |

# Reference Architecture

**Tier-3**
**Devices**

Drones

Video Cameras

Smartphones

AR/VR Devices

**Tier-2**
**Cloudlet**

WiFi or cellular

Low Latency
< 20 ms RTT
short tail

**Internet**

**Tier-1**
**Cloud**

aws

Google Cloud

OpenAI

ChatGPT

**Max 30 ms of headroom for "AI in the loop" @ 20 Hz**

**i.e., a 50 ms OODA loop ("Observe, Orient, Decide, Actuate")**

# Challenges & Headwinds

# Business Models

**Tier-1 and Tier-3 have well-established business models**

- **Tier-1: OPEX metered model (cloud as a utility)**
- **Tier-3: cost of each device includes amortized development cost**

**Tier-2 is new and more expensive: lower economies of scale**

**Premium pricing requires premium value**

**What is that value?**

# Cost/Benefit Tradeoffs

| | Tier-3 (Device) | Tier-2 (Cloudlet) | Tier-1 (Cloud) |
|---|---|---|---|
| **Deployment Cost** (with low-latency network if Tier-2 ) | ~ | – – | +++ |
| **System Administration Cost** (self-managed assumed free) | ~ | – – | +++ |
| **Software Development Effort** | – – | +++ | +++ |
| **Responsiveness** | +++ | ++ | – – |
| **Compute Capability** | – – – | ++ | +++ |
| **Elasticity** | ~ | – | +++ |
| **Privacy** | +++ | ++ | – |

***Cross-Layer Architecture Optimal***

# What is an Edge-Native Application?

**One that**

- **cannot run entirely at Tier-3** **(but needs Tier-3)**
- **cannot run entirely at Tier-1**
- **demands one or more unique *Tier-2 attributes***

  (Listed earlier: bandwidth scalability, low latency, sensor data privacy, high-availability, data sovereignty)

**Is one of these Tier-2 attributes essential for the use case?**

- **YES → Edge-Native application**
- **NO → Tier-2 is an expensive luxury**

  (Tier-1 + Tier-3 will always be cheaper, and adequate)

# Chicken-or-Egg Problem

**Software developers ask:** ***"Where is the infrastructure?"***

**Infrastructure deployers ask:** ***"Where are the edge-native applications?"***

***Breaking this deadlock is key to advancing Edge Computing***

# One Possible Solution

***Allow ad hoc creation and discovery of Tier-2 infrastructure***

- **rapid and easy deployment of Tier-2 infrastructure**
  **(even behind NAT firewalls)**
- **especially valuable in remote, under-provisioned locations**
- **make small-scale and incremental Tier-2 investments profitable**

**See technical report: "The Just-In-Time Cloudlet"**

Blakley, J., Meunier, M., Eiszler, T., Harkes, J., Satyanarayanan, M., CMU Technical Report, CMU-CS-23-138 October 20

**Dream: airline carry-on as JIT cloudlet**

Joe Speed

https://www.cmu.edu/scs/edgecomputing/news/fly-away-kit.html

32 to 192 core Ampere Cloud Native Processors for Vehicles, Robots, IoT, 5G and Space

1mo · Edited

NextComputing Ampere Fly-Away Kit rolling carry-on for portable 5G, cyber analytics, network forensics and data recording. Up to 256 Ampere CPU cores, 4TB RAM, 480TB NVMe. 5G small cell fits inside. https://lnkd.in/gkgvjWHV NextComputing engineering and Ampere power efficiency makes this possible.

# New Challenge:

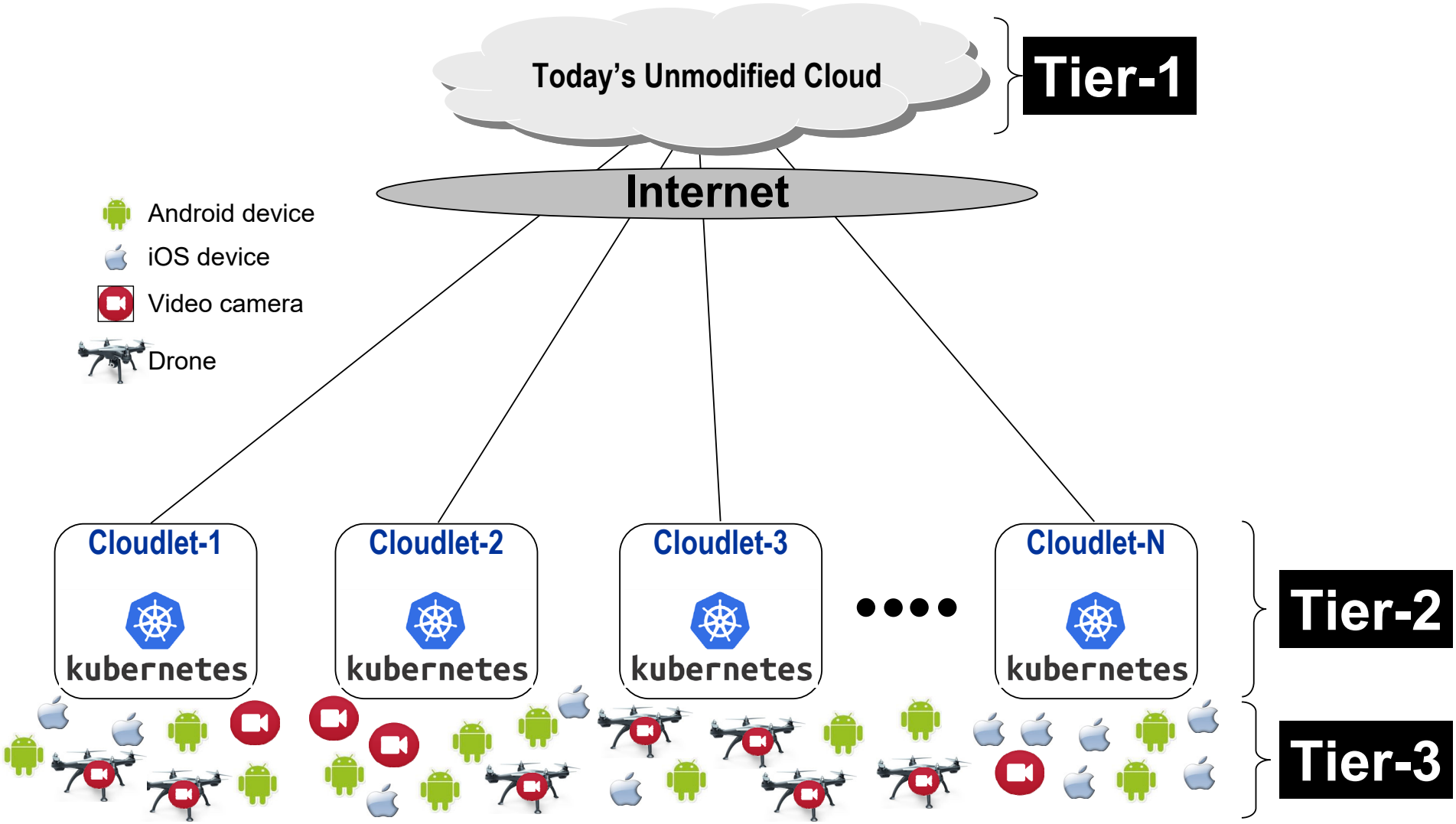
## *How to find the optimal cloudlet?*

# Problem

**Part 1:** ***How does a Tier-3 device pick the optimal cloudlet at Tier-2?***

**Part 2:** ***How can a diversity of cloudlets be supported?***

- ***telco-managed***
- ***hyperscaler-managed***
- ***just-in-time hyperconverged*** (ad hoc, possibly behind firewall)



# Observations

1. *Network Proximity is important*
   - weakly correlated with physical proximity
     200 km per millisecond in fiber at speed of light
   - physical proximity is neither sufficient nor necessary for network proximity
2. *"Motion-to-Photon Latency" (MTPL) is key*
   - complex end-to-end metric spanning compute and transmission
   - correlated with network latency and bandwidth
   - but also correlated with CPU speed/load, accelerators (e.g. GPUs), cache state, …
3. *Non-technical considerations matter*
   - business considerations
   - societal priorities (e.g. moral equivalents of handicapped parking)
   - legal constraints (e.g. GDPR privacy constraints on data placement)

# Sinfonia Requirements

1. *True end-to-end control*
   - edge-native applications are diverse, complex, and stateful
   - behavior and resource needs defy simple characterization relative to MTPL
   - new releases of an app may change its behavior and resource needs
   - decision logic best expressed as black-box code provided by developer

     Sinfonia should center discovery process around such black-box code

2. *Leverage emerging de facto standards*
   - Kubernetes widely used today at Tier-1
   - Prometheus resource monitoring within Kubernetes clusters

3. *Play nicely with emerging "walled gardens"*
   - proprietary mechanisms and ecosystems for edge computing

     (e.g., AlefEdge, Equinix, Nodeweaver, StackPath, …)
   - builtin decision-making for placement
   - developing for Sinfonia should allow easy use of those walled gardens

# Sinfonia Open Source Release

https://github.com/cmusatyalab/sinfonia

**For more details see paper:**

**"Sinfonia: Cross-tier orchestration for edge-native applications"**

Satyanarayanan, M., Harkes, J., Blakley, J., Meunier, M., Mohandoss, G., Friedt, K., Thulasi, A., Saxena, P., Barritt, B., Frontiers in Internet of Things, Volume 1, October 2022, https://doi.org/10.3389/friot.2022.1025247

# Steps in Sinfonia

# Code Components at all Tiers

*ST-1:* **Sinfonia code at Tier-1**

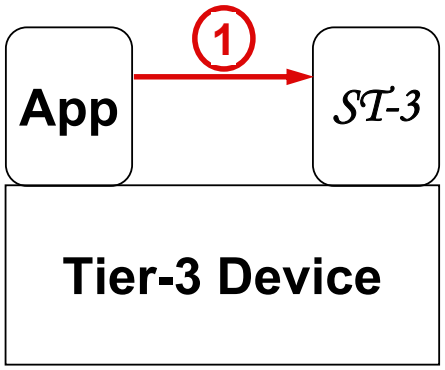
*ST-2:* **… Tier-2**

*ST-3:* **… Tier-3**

# Likely Scaling

(very rough, back-of-envelope estimates – easily off by one order of magnitude)

| | How Many | Why |
|---|---|---|
| *ST-1* (Tier-1, Cloud) | $10^2 – 10^3$ | Consistent with published estimate of roughly $10^5$ ISVs worldwide. Small ISVs likely to outsource their *ST-1* to shared registrar like NameCheap or GoDaddy. |
| *ST-2* (Tier-2, Cloudlets) | $10^5 – 10^6$ | **Reason-1:** inhabited area of the earth is $25 \times 10^6$ square miles; coverage area between 25 and 250 square miles per cloudlet.<br><br>**Reason-2:** ~$10^4$ cities worldwide and at least a few cloudlets per city + non-urban cloudlets<br><br>**Reason-3:** ~1% –10% of Tier-3 devices used for edge-native applications, with fan-in of $10^2$–$10^3$ per cloudlet |
| *ST-3* (Tier-3, Devices) | $10^9 – 10^{11}$ | Roughly 7 billion people on earth, each with 1 – 100 devices on average on-body or at home |




**App provides URL for root of orchestration**

- **URL is wired into source code** (or derived from environment, user, etc.)
- **root of orchestration is outside Sinfonia** (simple default possible)

**NULL URL**

- **“only use cloudlets discovered by ZeroConf”**
- **runtime config to force this also possible**

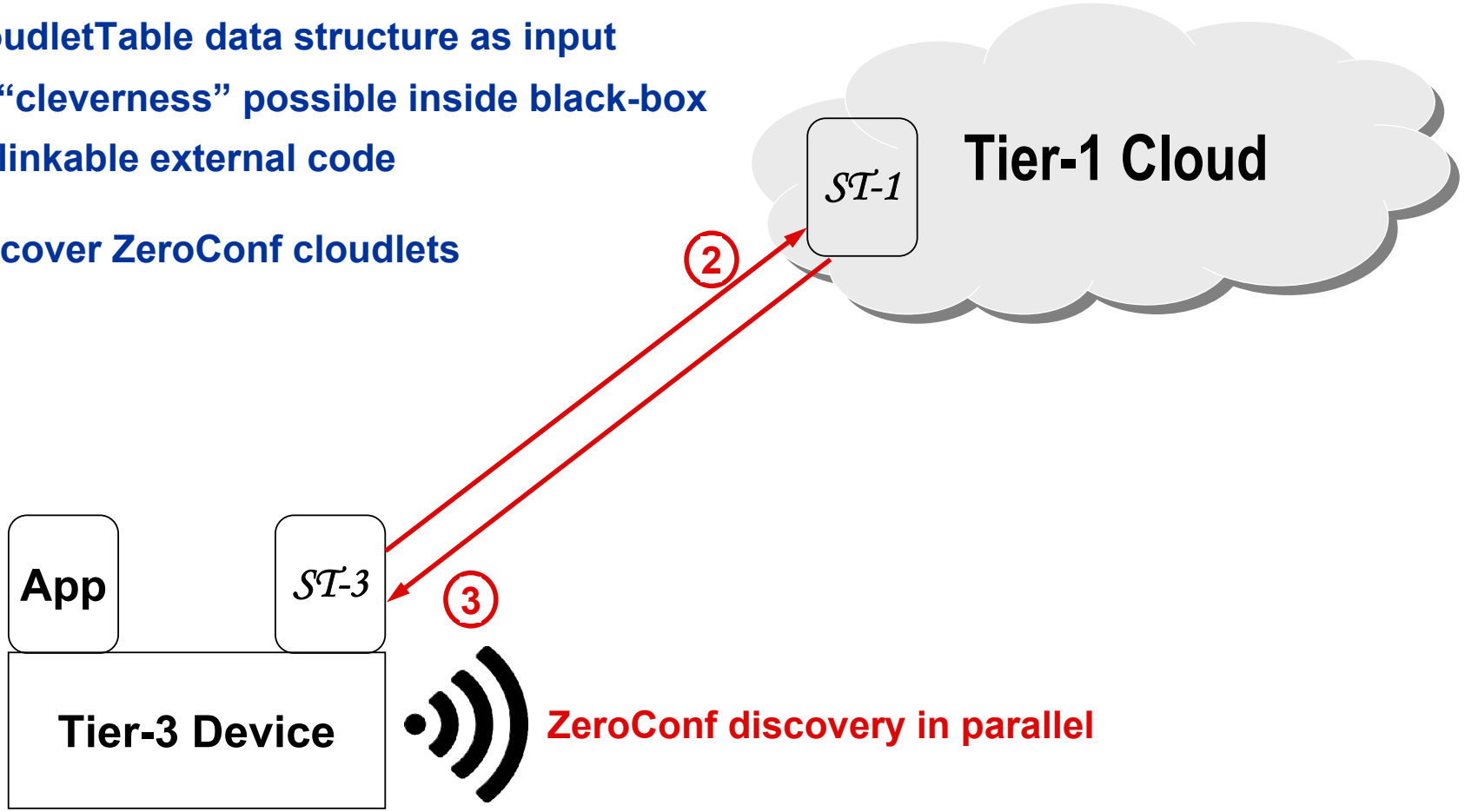
*ST-3* adds location info, device info, user info, etc.

*ST-1* invokes black-box code

*ST-1* provides CloudletTable data structure as input

- open-ended "cleverness" possible inside black-box
- dynamically linkable external code

Concurrently discover ZeroConf cloudlets



# CloudletTable at *ST-1*

**one row per cloudlet → $10^5$ – $10^6$ rows**

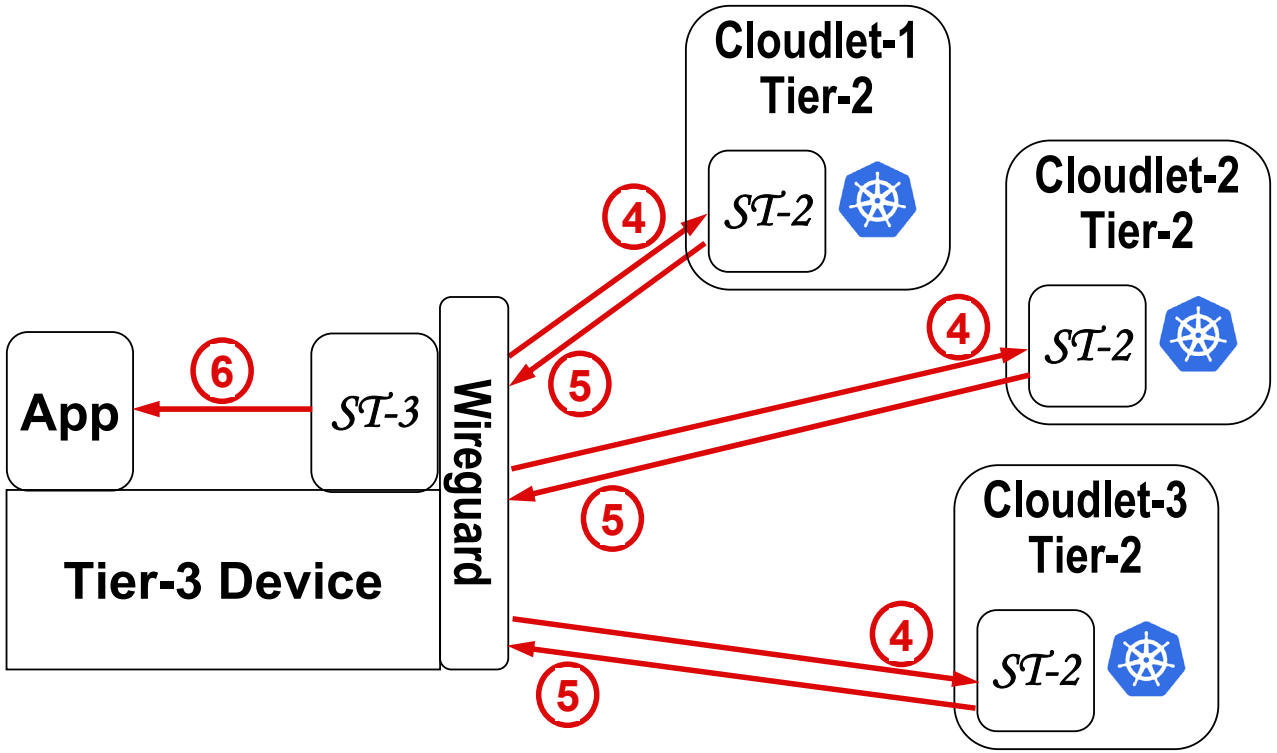
**Static Attributes** | **Dynamic Attributes**

| | UUID | Owner/ Operator | Hostname/ IP address | Geo-location | Public Key | Level of Trust | . . . | Fraction of Cores Utilized | Ping Time to ST1 | Ingress/Egress Bandwidth | . . . |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | |
| 2 | | | | | | | | | | | |
| 3 | | | | | | | | | | | |
| 4 | | | | | | | | | | | |

**Maintenance of CloudletTable is key responsibility of *ST-1***

**Open-ended list of attributes, likely to evolve**

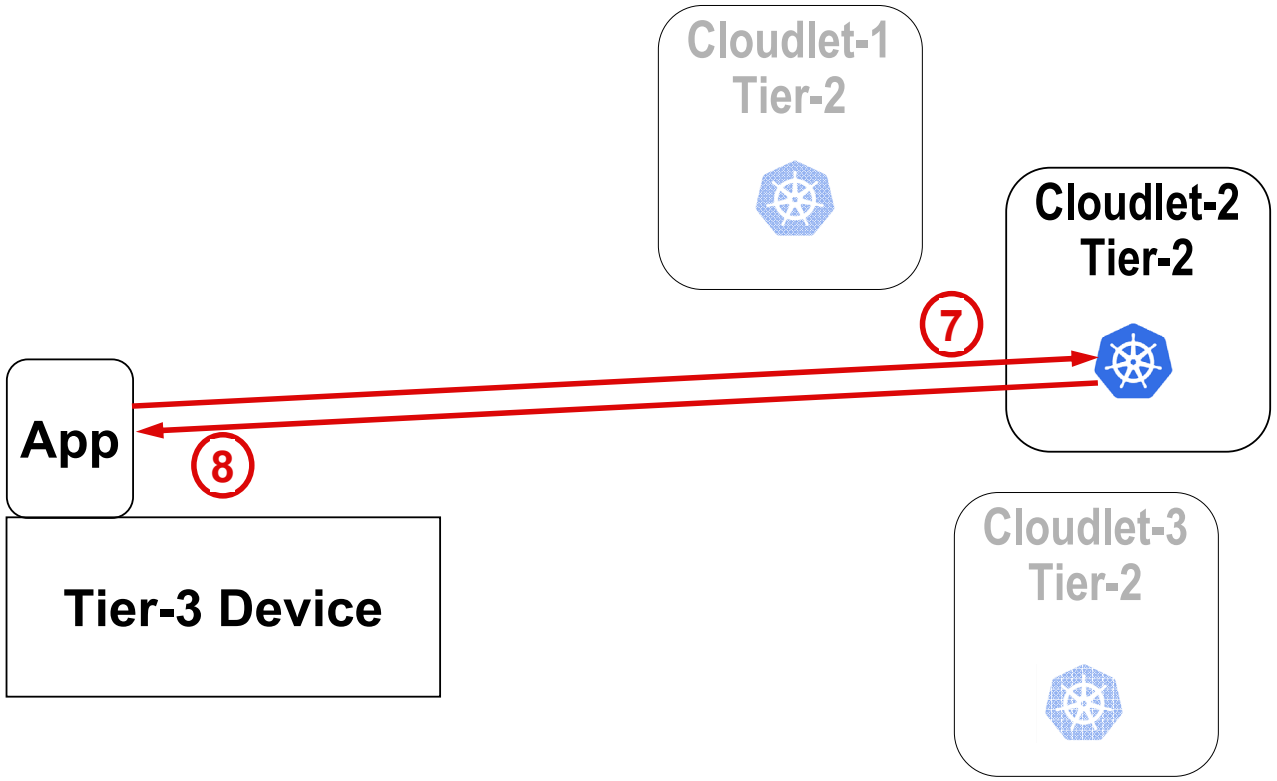
**Some standardization (ontology) of attributes needed**




**Short list of candidate cloudlets considered by *ST-3***

**Can be as short as just one cloudlet (e.g., only ZeroConf-discovered)**

**App given a chance to probe choices and indicate preference to *ST-3***




**Chosen back-end appears as *private network service***

# Registration of Cloudlets

**Anyone can offer Tier-2 services (competitive marketplace)**

***ST-2*** **reaches out to** ***ST-1*** **with offer and credentials**

- **up to** ***ST-1*** **to accept or not**
- **each** ***ST-1*** **and each** ***ST-2*** **choose these actions; no central control**

**Trust validation is responsibility of** ***ST-1*** **(e.g. via SGX, TrustZone, etc.)**

**Periodic keepalives from** ***ST-2*** **to** ***ST-1*** **with current attributes**

- **may include dynamic pricing details**

***Bottom-up and decentralized Tier-2 marketplace***

# How Will Edge Computing Transform Society?

# Focus on “Motion-to-Photon” Latency Systems

cyber-physical and cyber-human systems

***Ultralight Autonomous Drones***

mission-centric drone autonomy

***AR/VR with wireless offload***

especially Wearable Cognitive Assistance (AR+AI)

***EdgeVDI***

your legacy world anywhere anytime

# Maintaining Civil Assets is Expensive!

"Affectionately known as the "City of **Bridges**," **Pittsburgh** boasts 446 **bridges** – more than any other city in the world, including Venice, Italy."

**Large expense to local governments**

**Regular inspection would help ensure timely repairs**

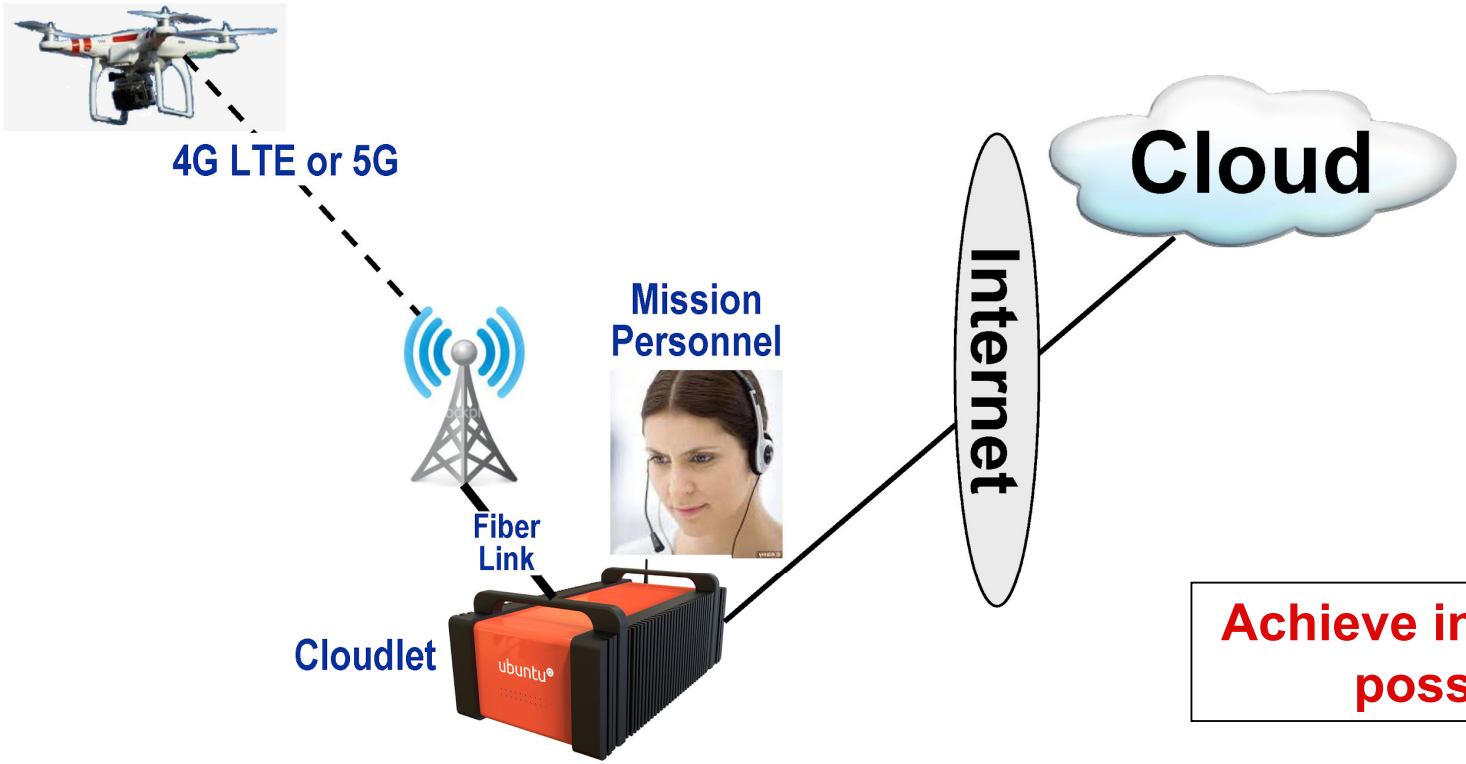
**Automating inspection is key**

**Our approach has enormous potential value for state and local governments nationwide**

***35,879 local governments in the US***

- **19,519 municipal governments**
- **16,360 town and township governments**
- **3,031 county governments**
- **(US Census Bureau)**

# Our Vision




**Benefits**

**Drone is small & lightweight**

**Powerful cloudlet nearby**

**Only 1-hop wireless latency**

**Drone has vast compute power without delay to cloud**

**Achieve in real-time what is only possible offline today**

# Ultralight Drone Autonomy for Active Vision

**Manual piloting limits safety and cost effectiveness**

**Autonomous drones increase safety by**

- **eliminating the need for workers to perform dangerous manual inspection tasks**
- **improving aviation safety by reducing likelihood for pilot error**

**Urban setting is especially challenging: many more hazards and obstacles**

***"Ultralight" → FAA regulations more flexible***

**below 250 grams total flight weight is magic**

# Pros & Cons of Edge Offload

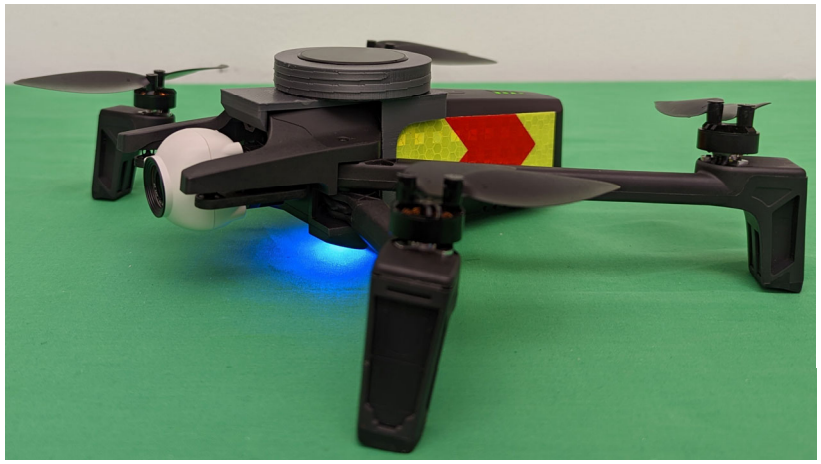
## Dual Use Technology

| | Military Context | Civilian Context |
|---|---|---|
| **Lightweight & Small** | + Carry in backpack<br>+ Small unit deployment<br>+ Low observability | + Public safety<br>+ Low FAA regulatory hurdle<br>(250 grams is magic) |
| **Cheap** | + Almost disposable hardware<br>+ Easy replacement | + Mass-market appeal<br>+ Faster ecosystem growth |
| **Weight-unlimited OODA Loop** | + Ability to "out-think" adversary<br>+ Easy AI escalation (arms race) | + Ease of software development<br>+ Faster development (devops) |
| **"Thin Client" Security** | + Drone capture does not compromise secrets | + Easier field maintenance |
| **Wireless dependence** | – Vulnerability to jamming<br>– Jam resistance hurts bandwidth | – Need for 5G coverage<br>– Need for edge computing |

# Our Initial Platform

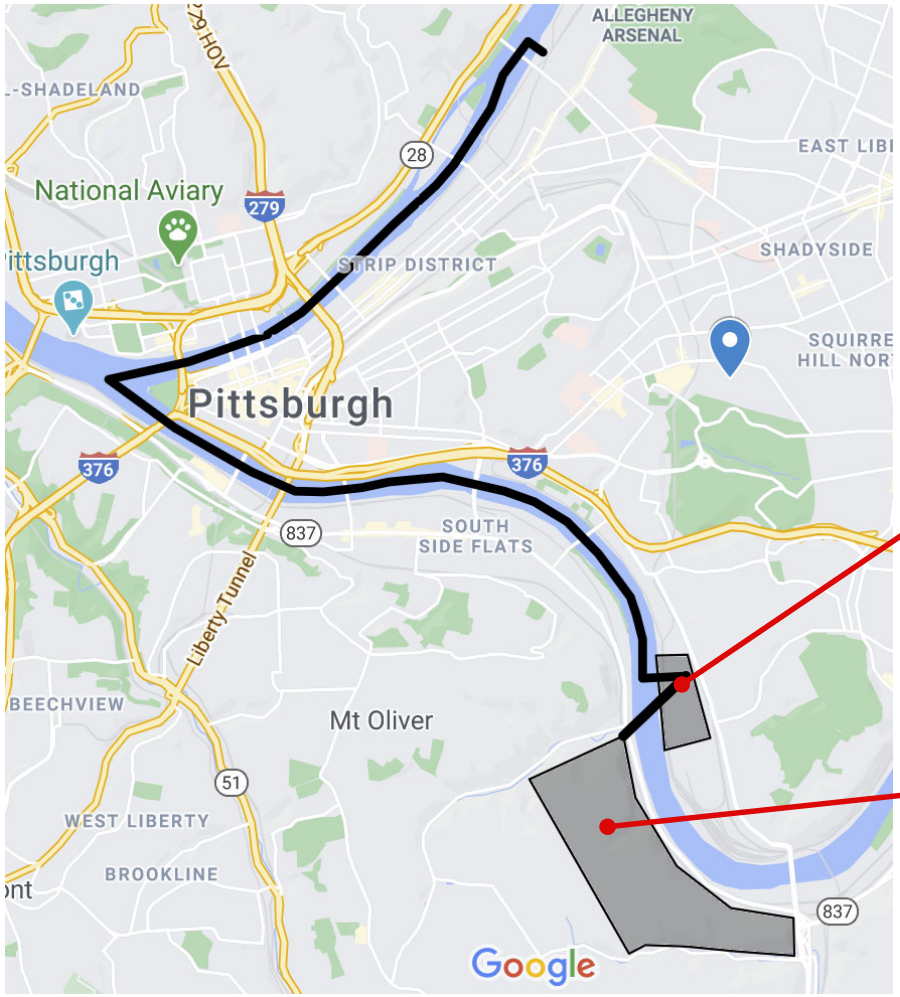
Parrot Anafi
320 g
$469 on Amazon
Max payload 50 g

Samsung Galaxy 4 Watch
26 g
$300 from Samsung



3D Printed Harness 14 g
Total Flight Weight 360 g

# Pittsburgh Drone Corridor




## Hazelwood Green

- **Industrial brownfield**
  **(former steel mill)**
- **Mill 19: CMU Advanced Manufacturing Facility**
- **Primary area of ops**

## Hays Wood

- **Wooded area**
- **Illegal dumping**
- **Invasive species**

# Area of Operations


**Mill 19 at Hazelwood Green**

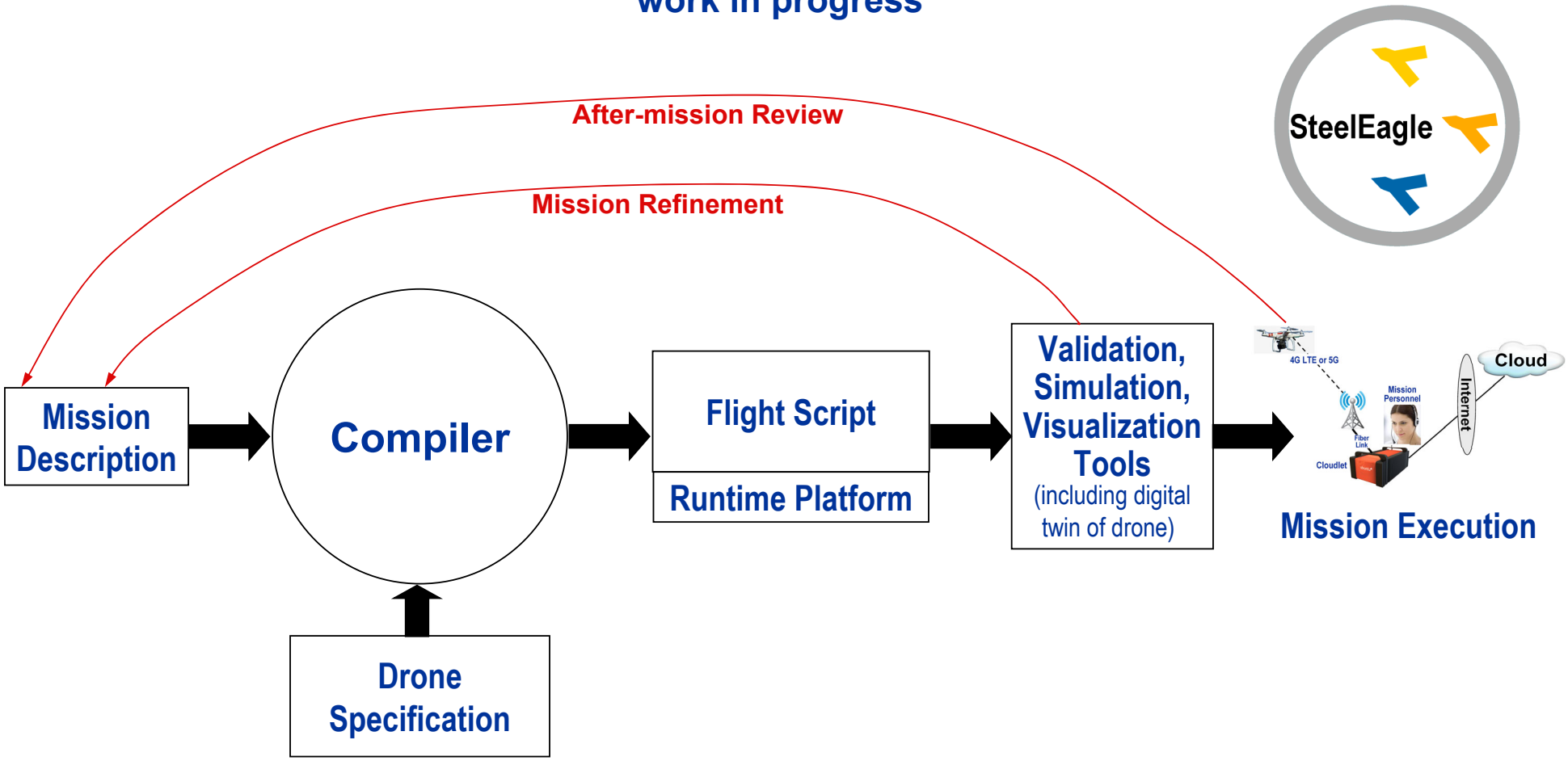
- **Southern terminus of flights**
- **Primary base of flight ops**


**NREC (National Robotics Engg. Center)**

- **Northern terminus of flights**
- **Secondary base of flight ops**

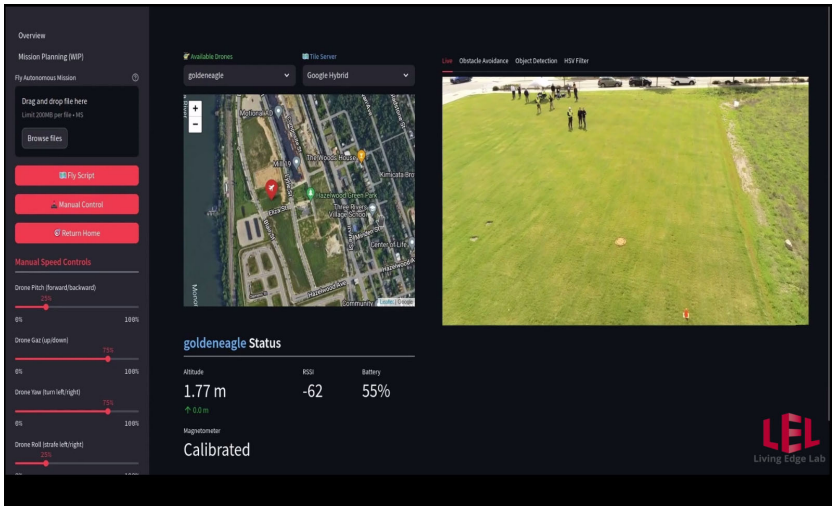
# Mission-Centric & Drone-Agnostic Tool Chain

work in progress




# Demo (April 24, 2024)

**Link to play video:** https://www.cmu.edu/scs/edgecomputing/videos/steeleagle_demo_04-04-24_short.mp4




# Closing Thoughts on Our Vision

*50 ms OODA loop for new cyber-human & cyber-physical AI applications*

**Humans are the standard against which AI is measured**

- **evolution took $10^9$ years to co-evolve humans, specialized neural circuits & apps**
- **we can't wait that long: ~10 years at most**
- **how to learn $10^8$ times faster than Nature?**

***Edge Computing is key***

- **enables real-time processing far beyond what could be worn or carried by humans**
- **larger, heavier, hotter, more energy-hungry, yet mobile**

**What is the role of cloud-based LLMs like ChatGPT ?**

see recent paper: **"Creating Edge AI from Cloud-based LLMs"** (https://dl.acm.org/doi/pdf/10.1145/3638550.3641126)

# Thank You!