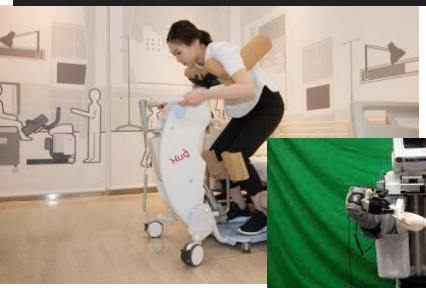
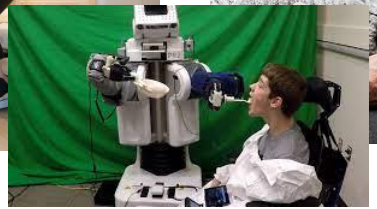
Minimal Mobile Systems via Cloud-based Adaptive Task Processing

Assistive Care

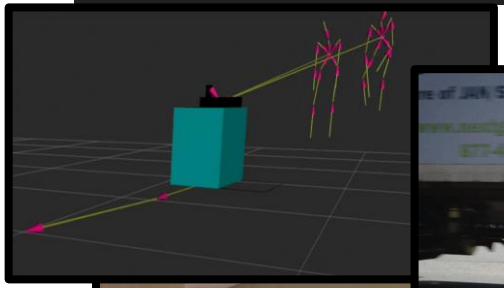


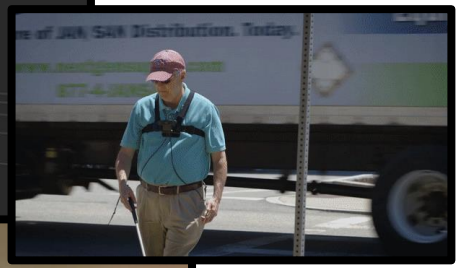


[Fuji Corporation]

[Park et al., 2020]

Personalized Mobility





Delivery

[Zipline, Dispatch]

Rehabilitation

[ReWalk]

Renato Mancuso

Sanja Arora

Hee Jae Kim

Lei Lai

Bassel Mabsout

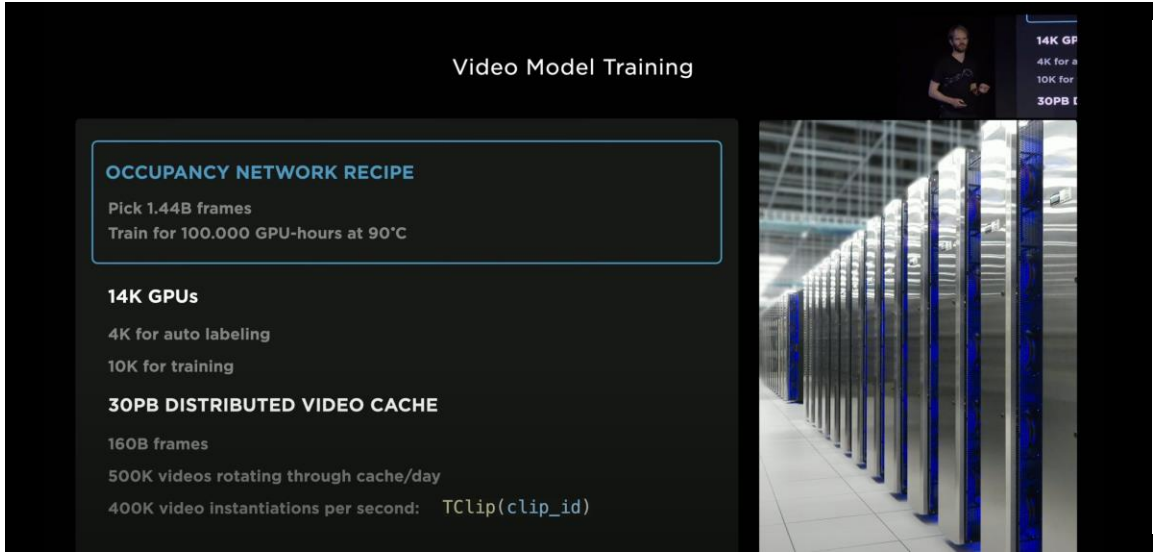
Eshed Ohn-Bar

Boston University

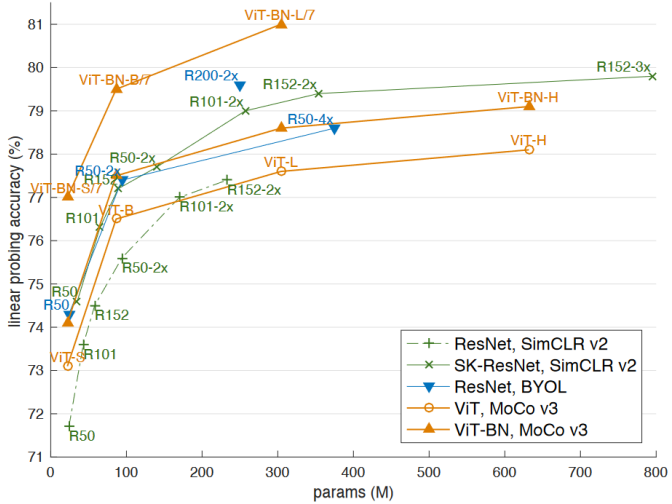
Minimal Mobile Systems via Cloud-based Adaptive Task Processing

Cost	$1,000	$15,000	$100,000
Battery	30 min	30 min	90 min

Minimal Mobile Systems via Cloud-based Adaptive Task Processing



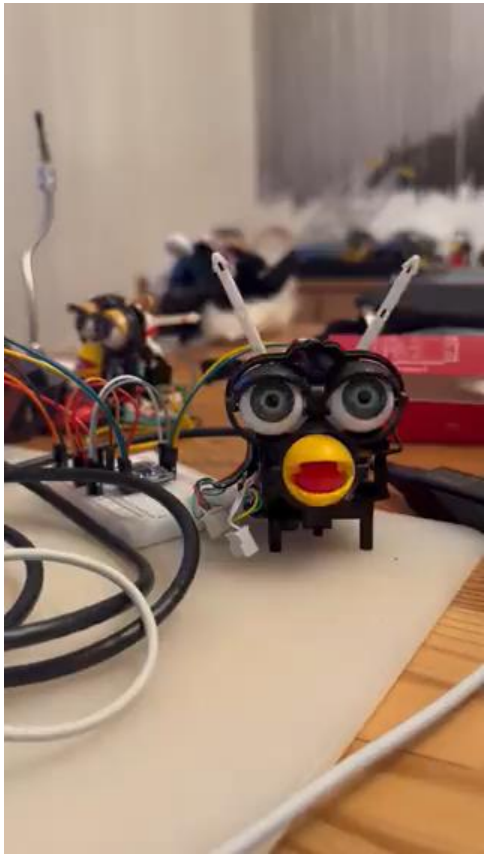
[Tesla AI Day, 2022]



[Chen et al, 2022]

Cloud-based Inference?

[BeMyEyes, OpenAI]



[@jessicacard]

Edge Device DNN Inference is Costly

Category	IoT/Edge Devices	GPU-Based Edge Devices		Custom-ASIC Edge Accelerators		FPGA Based	CPU	HPC Platforms GPU		
Platform	Raspberry Pi 3B [34]*	Jetson TX2 [69]	Jetson Nano [36]	EdgeTPU [35]	Movidius NCS [37]♦	PYNQ-Z1 [64]	Xeon	RTX 2080	GTX Titan X	Titan Xp
CPU	4-core Ctx.A53 @1.2 GHz*	4-core Ctx.A57 2-core Denver2 @2 GHz	4-core Ctx.A57 @1.43 GHz	4-core Ctx.A53 & Ctx.-M4 @1.5 GHz	N/Ap	4-core Ctx.A9 @650 MHz	2x 22-core E5-2696 v4 @2.20GHz	N/Ap*	N/Ap	N/Ap
GPU	No GPGPU	256-core Pascal μA	128-core Maxwell μA	N/Ap	N/Ap	N/Ap	N/Ap	2944-core Turing μA	3072-core Maxwell μA	3840-core Pascal μA
Accelerator	N/Ap	N/Ap	N/Ap	EdgeTPU	Myriad 2 VPU	ZYNQ XC7Z020	N/Ap	N/Ap	N/Ap	N/Ap
Memory†	1 GB LPDDR2	8 GB LPDDR4	4 GB LPDDR4	N/Av*	N/Av	630 KB BRAM 512 MB DDR3	264 GB DDR4	8 GB GDDR6	12 GB GDDR5	12 GB GDDR5X
Idle Power‡	1.33	1.90	1.25	3.24	0.36	2.65	≈70	≈39	≈15	≈55
Average Power‡	2.73	9.65	4.58	4.14	1.52	5.24	300 TDP	≈	≈100	≈
Platform	All	All	All	TFLite	NCSDK	TVM/FINN	All	All	All	All

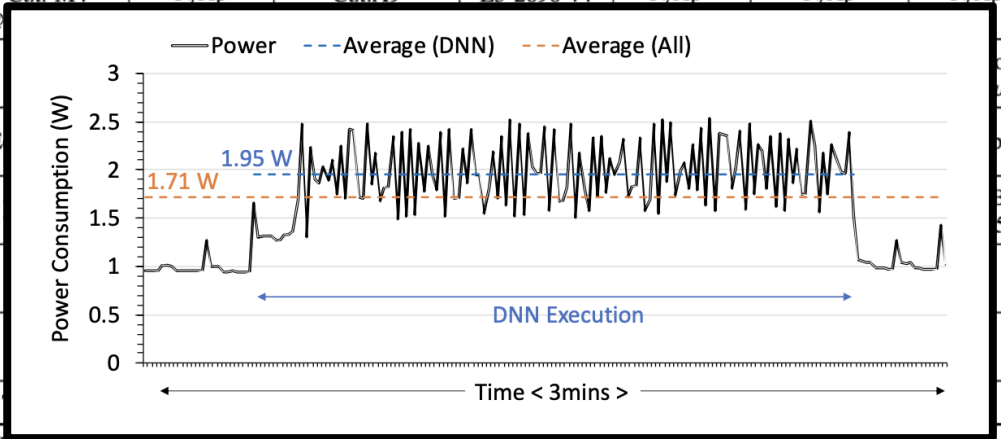
† Effective memory size used for acceleration/execution of DNNs, e.g., GPU/CPU/Accelerator memory size. * Ctx.: Arm Cortex. N/Ap: Not applicable. N/Av: Not available.
‡ : Measured idle and average power while executing DNNs, in Watts. ⋆ : Raspberry Pi 4B [70], with 4-core Ctx.A72 and maximum of 4 GB LPDDR4, was released after this paper acceptance. With better memory technology and out-of-order execution, Raspberry Pi 4B is expected to perform better. ♦ Intel Neural Compute Stick 2 [61] with a new VPU chip and support for several frameworks was announced during paper submission, but the product was not released.

[Hadidi et al, Characterizing the Deployment of Deep Neural Networks on Commercial Edge Devices, IISWC, 2019]

Edge Device DNN Inference is Costly

Category	IoT/Edge Devices	GPU-Based Edge Devices		Custom-ASIC Edge Accelerators		FPGA Based	CPU	HPC Platforms GPU		
Platform	Raspberry Pi 3B [34]*	Jetson TX2 [69]	Jetson Nano [36]	EdgeTPU [35]	Movidius NCS [37]♦	PYNQ-Z1 [64]	Xeon	RTX 2080	GTX Titan X	Titan Xp
CPU	4-core Ctx.A53 @1.2 GHz*	4-core Ctx.A57 2-core Denver2 @2 GHz	4-core Ctx.A57 @1.43 GHz	4-core Ctx.A53 & Ctx.-M4 @	N/Ap	4-core Ctx.A9	2x 22-core E5-2696 v4	N/Ap*	N/Ap	N/Ap
GPU	No GPGPU	256-core Pascal μA	128-core Maxwell μA							ore μA
Accelerator	N/Ap	N/Ap	N/Ap	E						
Memory†	1 GB LPDDR2	8 GB LPDDR4	4 GB LPDDR4							5X
Idle Power‡	1.33	1.90	1.25							
Average Power‡	2.73	9.65	4.58							
Platform	All	All	All							

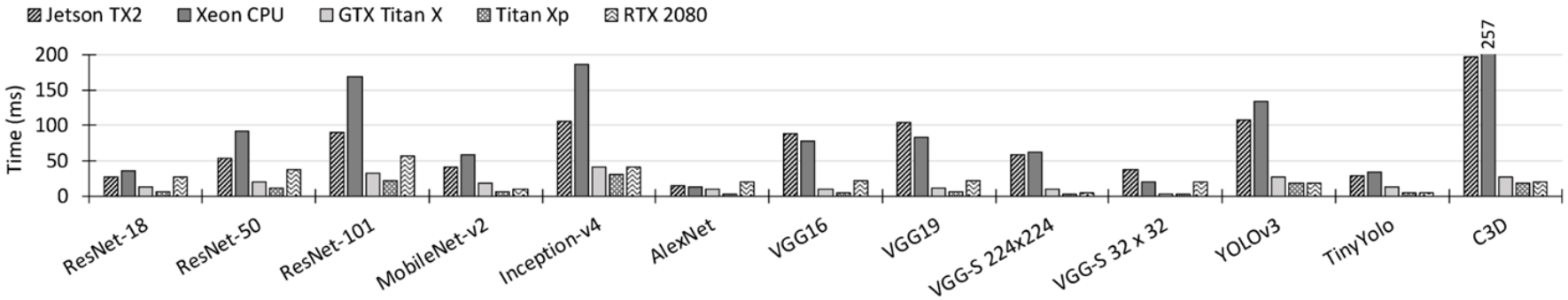
† Effective memory size used for acceleration/execution of D
‡ : Measured idle and average power while executing DNNs,
acceptance. With better memory technology and out-of-orde
and support for several frameworks was announced during p

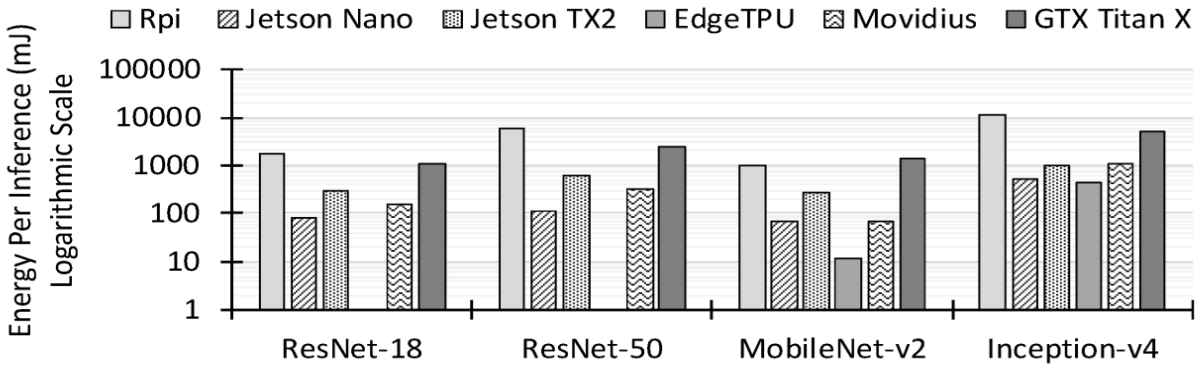


iRobot: 25-50+% Power increase in DNN Inference

[Hadidi et al, 2019]

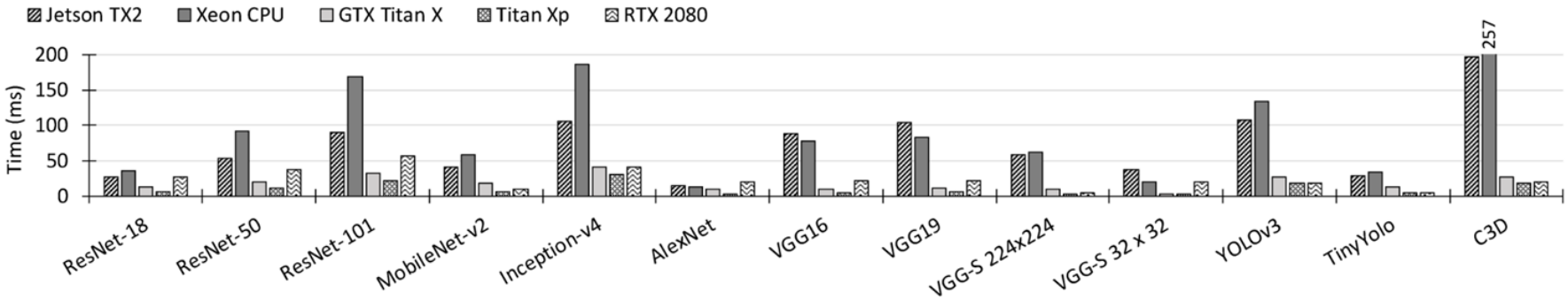
Edge Device DNN Inference is Costly



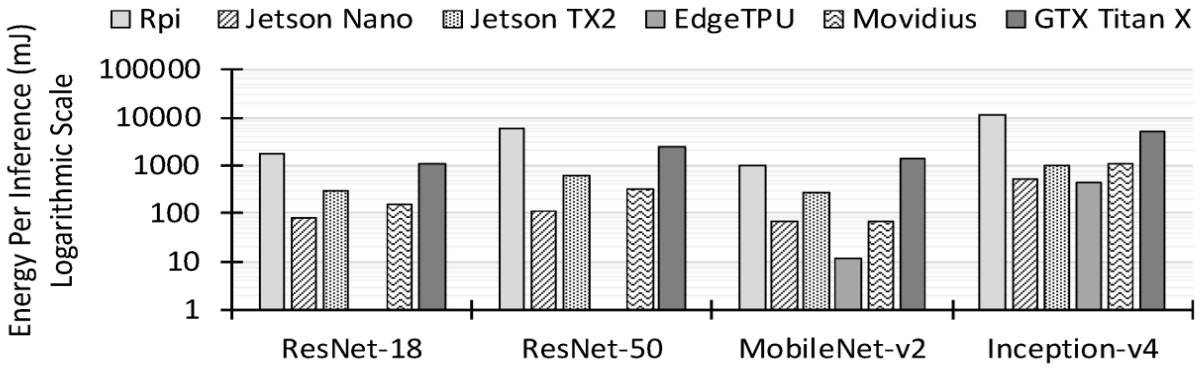


[Hadidi et al, 2019]

Edge Device DNN Inference is Costly

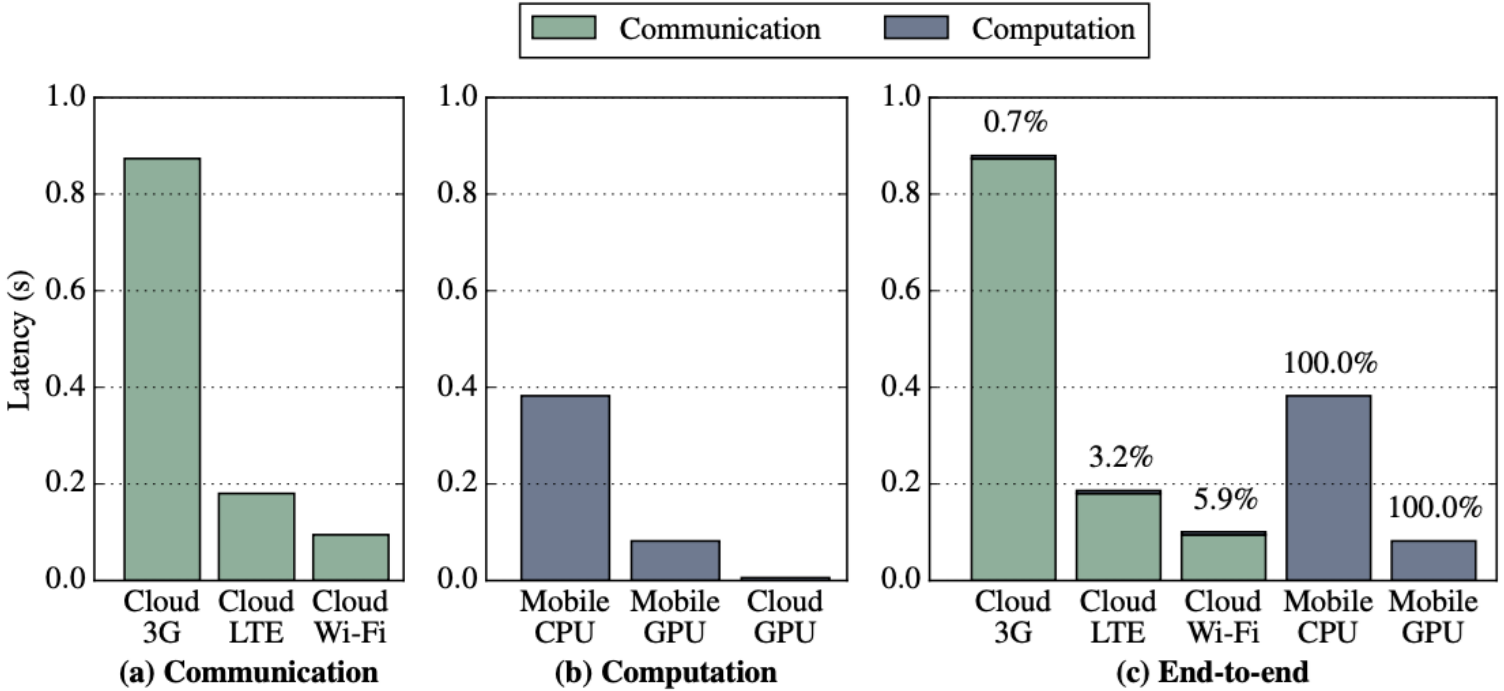


	mIOU	Params	MAdds
MNet V2*	**75.32**	**2.11M**	**2.75B**
	77.33	2.11M	152.6B
ResNet-101	80.49	58.16M	81.0B
	82.70	58.16M	4870.6B



[Hadidi et al, 2019]

Communication Latency is Decreasing Rapidly



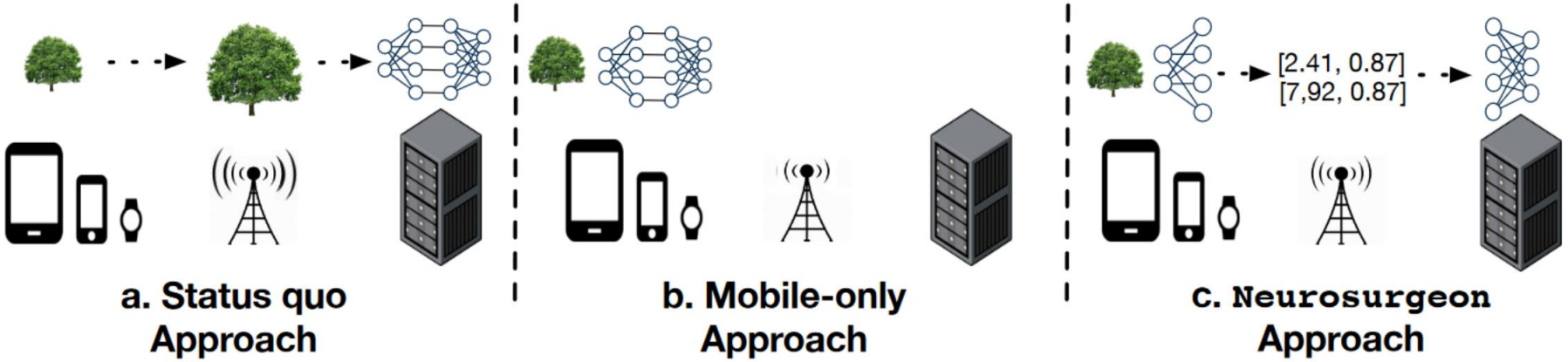
Transmitting Images

Execution

Total

[Kang et al, ASPLOS 2017]

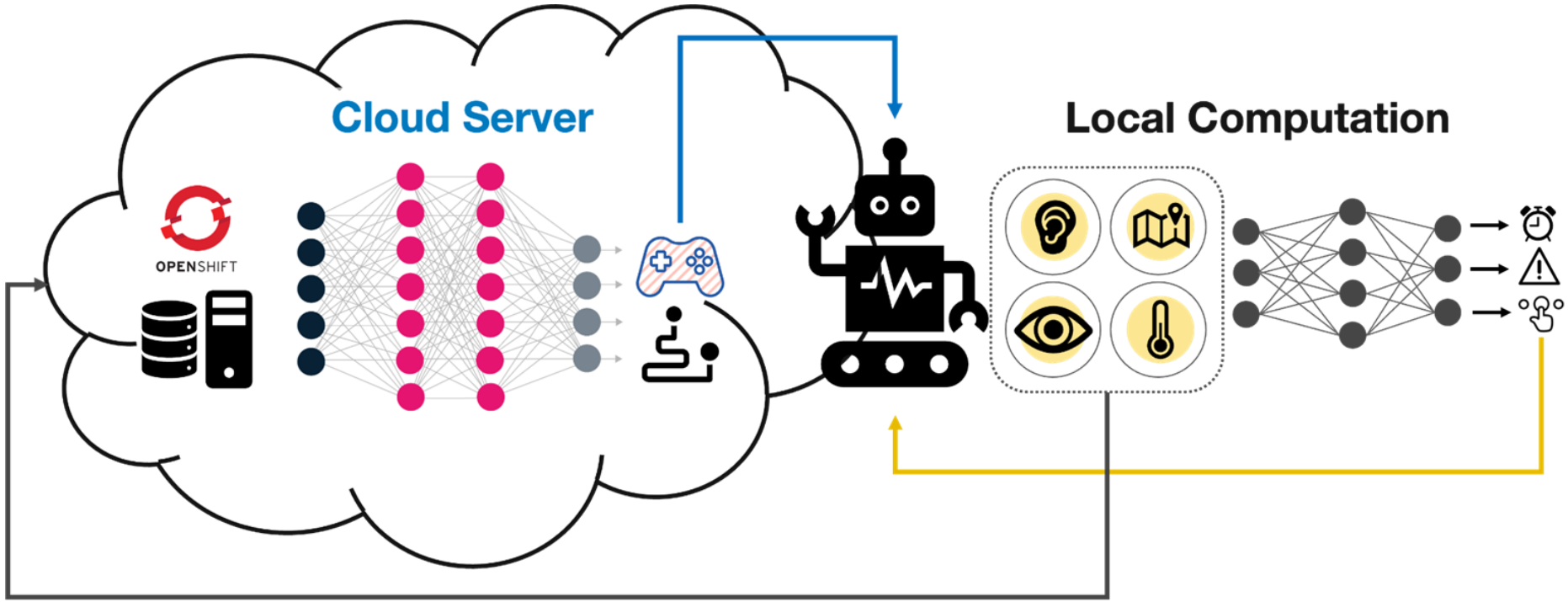
Prior Work - Example



- May still have unacceptable latency
- Not ***task nor context-dependent***
- Cloud can run bigger and better models

[Kang et al, ASPLOS 2017]

How to Combine the Best of Both Worlds?

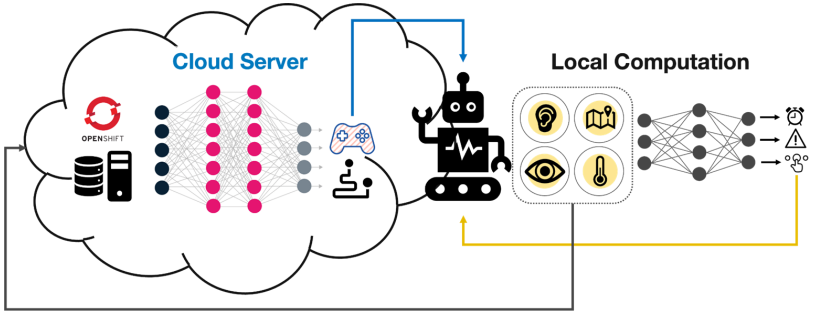


Task-Aware Markov Decision Process Formulation

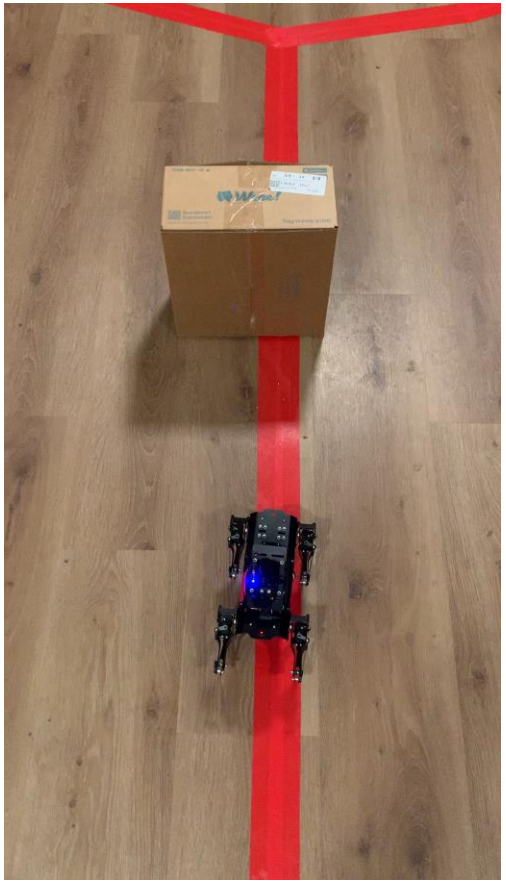
- **States**: Data coming from the robot (e.g., ***sensor data***, battery state, safety), potentially cloud server state (e.g., queues, prioritization, running tasks).

- **Actions**: The agent must select ***the task*** to perform and whether ***locally or on the cloud***.

- **Task Reward**: Collision avoidance, path adherence, social disruption.

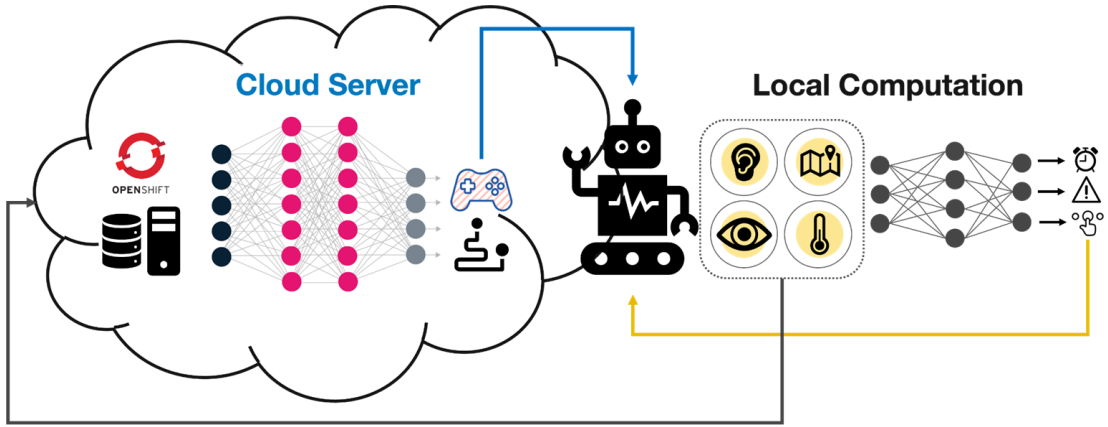


Preliminary Results



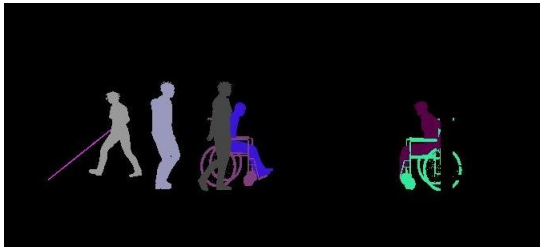
Open Challenges

- Standardized benchmarking?
- Latency model?
- Cloud processes?
- Local adaptation?



Standardized Benchmark with Streaming Perception

Accessibility, Vision, and Autonomy Challenge @ CVPR 2023



Win **$500** to detect pedestrians and mobility aids:
accessibility-cv.github.io

Toyota suspends all self-driving vehicles at Paralympic Games after collision with athlete

Scooter Doll - Aug. 27th 2021 10:58 am PT

Pitt suspends delivery robots after wheelchair user reports safety hazard

JAMIE MARTINES | Thursday, Oct. 24, 2019 7:27 p.m.

SUPPORT LOCAL JOURNALISM