Serving Machine Learning from the Edge

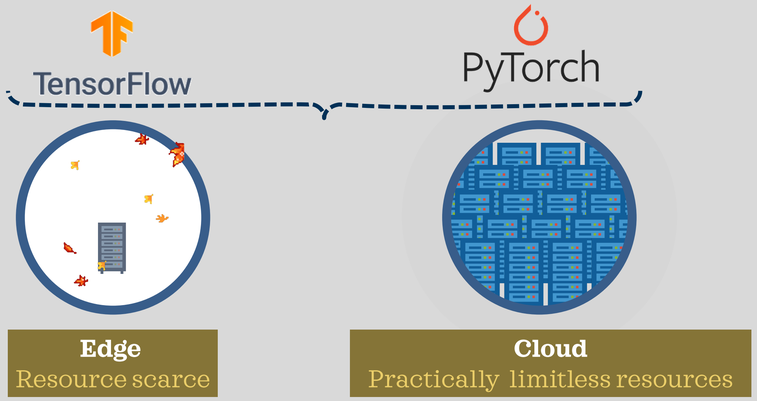
Misun Park, Ketan Bhardwaj, Ada Gavrilovska

School of Computer Science

Georgia Tech | Center for Research into Novel Computing Hierarchies

Problem Statement

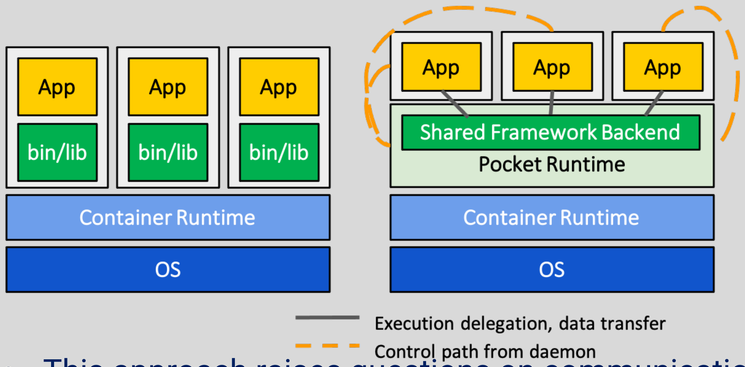
- Resource-scarce edge vs. Resource abundant cloud



- A growing demand to serve a variety of complex applications, including video analytics and data analytics, at the edge
- Two contradictory goals: tight computing resource budget on edge servers and complex and resource-demanding applications contradict each other

Pocket Approach

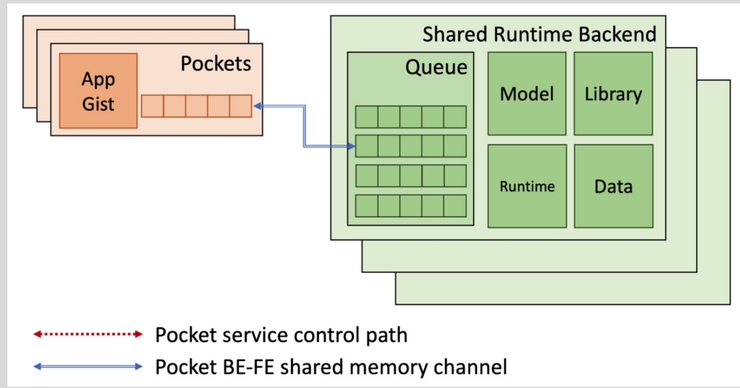
- Intuitive solution: having a shared runtime backend for multiple application front ends



- This approach raises questions on communication channel and resource management

Design Details

- Private queues for each application frontend and right isolation mechanism is implemented

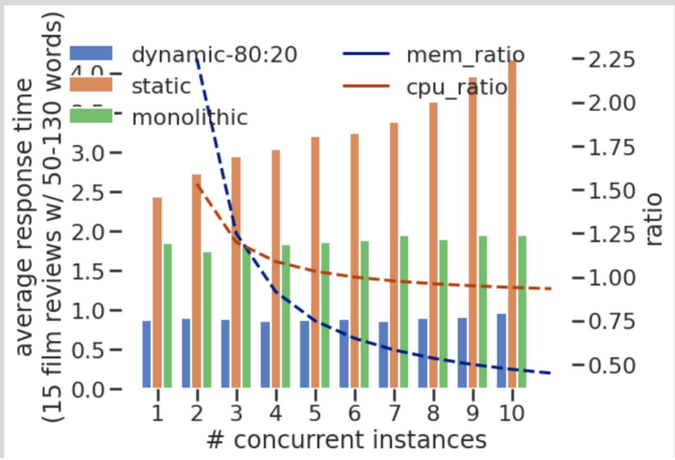


- Proper resource is managed according to its policy

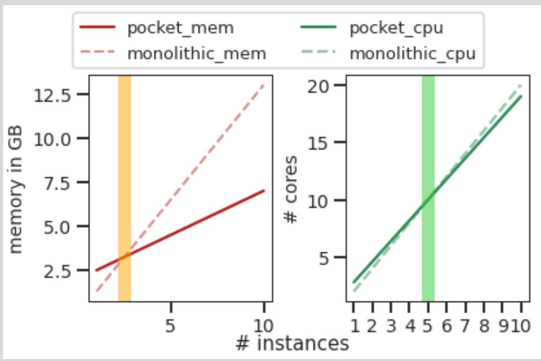
resource type	granularity	resource realloc	amount to transfer
CPU core	connection	dynamic	(0, 1)
memory	function	static	N/A

Evaluation

- Better resource efficiency & performance
  Pocket serves the same number of applications with less amount of resource, while showing comparable performance (varying in applications)



- Better scalability
  Overall, with the same amount of resource, Pocket servers a greater number of concurrent application instances at once



Conclusion

- Transplanting cloud-native technologies to edge does not work in terms of performance and scalability, because edge is resource scarce but the workloads on demand are heavily resource consuming.
- Pocket was started from intuitive idea that separating applications into (1) common part and (2) app-specific part, and make the common part shared
- That intuitive idea raises several design challenges and Pocket tackles it with its unique design, suggesting how this type of shared backend should be designed implemented
- Pocket minimizes the change app deployment and packaging so that the benefits from the large developer community won't disappear