Risk-aware Adaptive Distributed Computation Placement on the Tactical Edge

A. Ali-Eldin, P Shenoy (UMass), B. Jalaian (ARL)

Relevant to IOBT CRA Research Area / Task: 2.2

Addressed Army Need: Computation Adaptation in a dynamic battlefield with network dynamics.

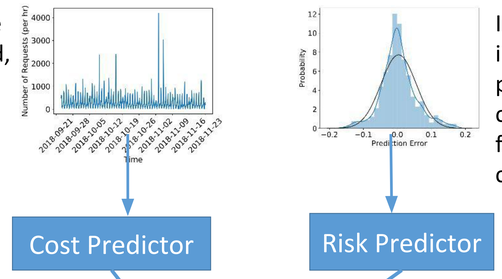
Challenge

- How to incorporate future knowledge of the battlefield risks and dynamics in adaptive placement of computations on the IoBT tactical edge
- The tactical edge has limited heterogeneous resources, where all computations in the battlefield compete to run.
- Computations can fail temporarily due to, for example, network dynamics or battery depletion

Approach

- Use two predictors, a **risk** predictor and a **cost** predictor that use historical knowledge about the risk/cost, plus any expert knowledge to predict the future risk/cost.
- Use **Multi-period Portfolio Optimization (MPO)** to decide where to place computations in the tactical edge, integrating the prediction output
- Each edge server has **cost** and **risk** to host a computation, e.g., Humvee with WIN-T with servers, V-bats carrying pico-clusters, etc

Inputs can include predicted demand, battery drain, communication costs, etc



Inputs can include hit probability, communication failures, risk of compromise, etc

Formulation

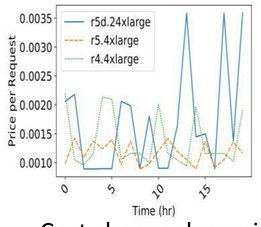
- Solve the following (convex) optimization problem for *H* look-ahead predicted time units
- Maximize risk-adjusted Savings

$$\text{Maximize} \sum_{\tau=t}^{t+H-1} E\left[Return(\tau)\right] - \left(cost(\tau) + \alpha(Risk(\tau))\right)$$

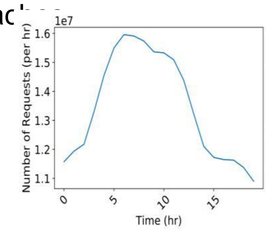
- The optimization needs to be convex

Results

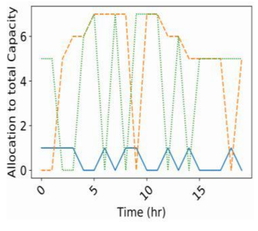
- The technique was tested for transient servers running latency-critical applications.
- A transient servers can **fail at any point**
- The cost of servers is dynamic in an open market
- For all our experiments, using our technique has resulted in no loss of computations after failures + a reduction in the cost of running up to 50% compared to other state-of-the-art approac



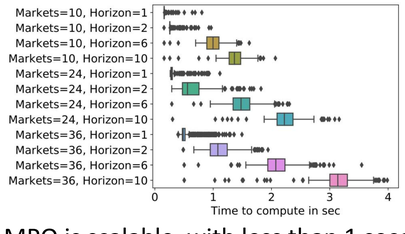
Cost change dynamics for three markets



Workload Dynamics



MPO output



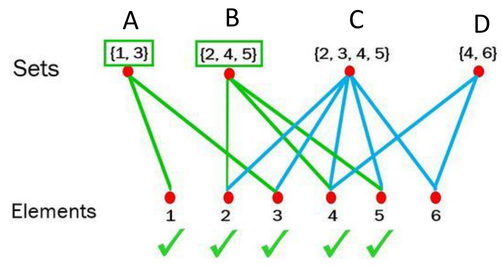
MPO is scalable, with less than 1 second computation time for large number of resource types, and a moderate horizon length

Conclusions

- Multi-period portfolio optimization is a promising technique to be used in the IoBT context
- It allows the commander to have a mechanism for risk-adjusted computation placement in the battlefield - the risk tolerance and cost guide the placement

Ongoing and Future Work

- Collaboration with Task 2.1 on the integration of their work on sub-modularity with multi-period portfolio optimization
  - Gives the commander the ability to set affinity rules for sets of resources e.g., Computations should include at least one resource from sets A, B, C, and D taking into account the risk
- Working on better predictors for the states in the battlefield
- Deployment and Experimental Validation of adaptive placement of ML inference computations on resource-constrained edge.



Publications and Software

- Ali-Eldin, A., Westin, J., Wang, B., Sharma, P., & Shenoy, P. (2019, June). Spotweb: Running latency-sensitive distributed web services on transient cloud servers. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*
- **Best Paper Runner-up** at ACM HPDC 2019
- Software available: http://github.com/umassos/spotweb-hpdc19

POINT OF CONTACT:

Prashant Shenoy
shenoy@cs.umass.edu