Corralling the Computing Continuum:
Mobilizing Modern Distributed Resources for Machine Learning and Accessible Computing

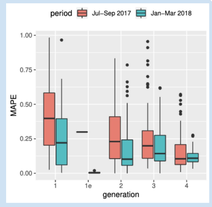
Matt Baughman*, Ian Foster*‡ (Advisor) & Kyle Chard*‡ (Advisor)
University of Chicago* & Argonne National Lab‡

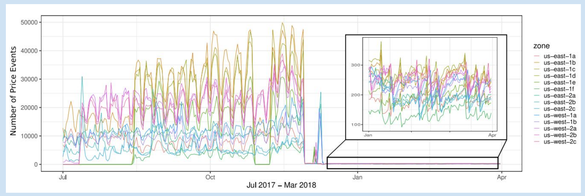
The Idea Behind the Journey

- The Computing Continuum
  - Moving compute and data where it is best
  - Computing remotely is as simple and locally
- Challenges
  - Existing systems focus on specific use cases or people
  - Keeping humans in the optimization loop
  - Demonstrating “compute anywhere” principles in real use cases
- Solution
  - Build on existing infrastructure
  - Remove the tough choices from people while leaving enough control for what is necessary
  - Demonstrate usefulness for distributed machine learning through serverless FL
- Corralling the Continuum can be made easy

Cost Prediction

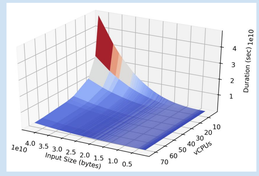
- Computing costs change short and long term
- If we understand these changes, we can better predict costs for cloud-based workflows
- Prediction of preemption also allows us to use less reliable resources, driving down costs further
- We demonstrate less than 4% error for expected cost given a workload duration

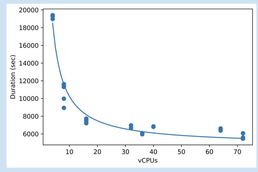




Workload Profiling

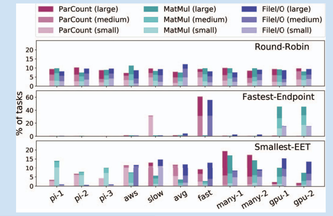
- Given we can predict compute costs given duration, we need to predict workload duration
- Create simple, composable models by exploring over a single variable (e.g., input size, RAM, etc.)
- Aggregate these models into a composite model
- Selectively run experiments to provide new data to refine these composite models
- Demonstrate a ~30% error for composite models with a single retraining point
- Automated experimental design allows predictions as accurate in 6 selected experiments as 14 random experiments





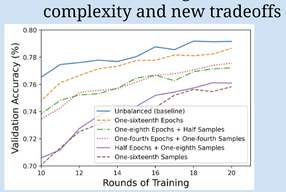
Putting it Together

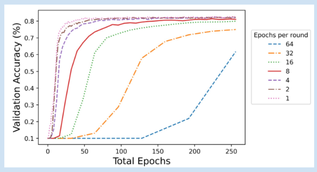
- DELTA
  - Distributed Execution of Lambdas with Tradeoff Analysis
  - Incorporating the profiling and prediction with provisioning
- Placing serverless tasks on heterogeneous compute with notions of cost and execution characteristics

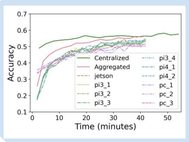


Trade-Offs and Optimizations in Serverless FL

- FL is difficult
  - Hyperparameter selection
  - Model configuration
  - Data wrangling
- Abstract away the difficult portions from the user to the system
  - Serverless is the solution
- Coordinating amongst more resources means greater complexity and new tradeoffs
- Demonstrate predictable learning characteristics with respect to federated learning workflow hyperparameters
- Identified repeatable patterns in the experiments
- Use those patterns to automate away the configuration from end-users

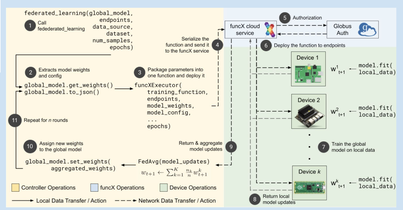


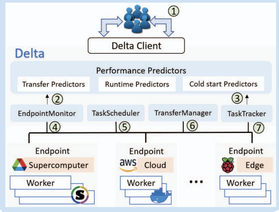




Serverless-Based Federated Learning

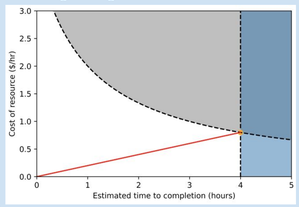
- FLoX
  - Federated Learning on funcX
- Pushing AI training to where the data is located using serverless
- Demonstrates capability of incorporating extremely low power resources alongside cloud and HPC

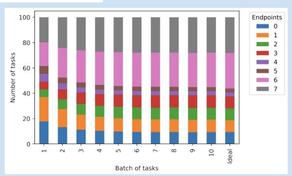




Putting it Together… Pt. 2

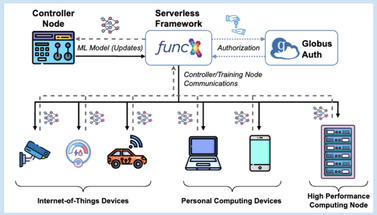
- We can use the same placement mechanism from DELTA to set hyperparameters on an endpoint by endpoint basis
- End-user configurability can be as transparent as desired
  - More science less system tuning
- Automatically incorporate diverse data sources and compute resources for easily configurable, multi-site federated learning using serverless





Ongoing and Future Work

- Integrate DELTA and FLoX
  - With portable data and compute, why not move both?
  - Automatically push compute to the data or data to compute based on user-specified optimization criteria and available resources
  - Explore the possibility of using FL to improve DELTA’s performance estimates, thereby also accelerating the FL process itself
- Incorporate traditional distributed training capabilities into FLoX to perform FL across multiple HPC resources
- Explore and incorporate methods for automating experiment workflows and resource configuration
- Extend DELTA to include support for complex workflows rather than only bag-of-task type workloads
- Demonstrate significantly greater scalability across multiple large machines simultaneously as well as diverse edge hardware
  - And applications to domain science problems



Relevant Works/Bibliography

- Matt Baughman, Nathaniel Hudson, Ian Foster, Kyle Chard. “Balancing Federated Learning Trade-Offs for Heterogeneous Environments.” 2023. PerCom 2023, WiP.
- Simon Caton, Matt Baughman, Christian Haas, Ryan Chard, Ian Foster, Kyle Chard. “Assessing the Current State of AWS Spot Market Forecastability.” 2022. SuperCompCloud Workshop at SC.
- Matt Baughman, Ian Foster, Kyle Chard. “Exploring Tradeoffs in Federated Learning on Serverless Computing Architectures.” 2022. eScience.
- Nikita Kotsehub, Matt Baughman, Ryan Chard, Nathaniel Hudson, Panos Patros, Omer Rana, Ian Foster, Kyle Chard. “FloX: Federated Learning with FaaS at the Edge.” 2022. eScience.
- Matt Baughman, Ian Foster, Kyle Chard. “Enhancing Automated FaaS with Cost-aware Provisioning of Cloud Resources.” 2021. eScience.
- Rohan Kumar, Matt Baughman, Ryan Chard, Zhuozhao Li, Yadu Babuji, Ian Foster, Kyle Chard. “Coding the Computing Continuum: Fluid Function Execution in Heterogeneous Computing Environments.” 2021. Heterogeneous Computing Workshop at IPDPS.
- Matt Baughman, Rohan Kumar, Ian Foster, Kyle Chard. “Expanding Cost-Aware Function Execution with Multidimensional Notions of Cost.” 2020. High Performance Serverless Computing Workshop at HPDC.
- Matt Baughman, Nifesh Chakubaji, Hong-Linh Truong, Krists Kreics, Kyle Chard, Ian Foster. “Measuring, Quantifying, and Predicting the Cost-Accuracy Tradeoff.” Benchmarking, Performance Tuning, and Optimization for Big Data Applications Workshop at BigData.
- Matt Baughman, Simon Caton, Christian Haas, Ryan Chard, Rich Wolski, Ian Foster, Kyle Chard. “Deconstructing the 2017 Changes to AWS Spot Market Pricing.” 2019. ScienceCloud Workshop at HPDC.

Contact: mbaughman@uchicago.edu