Geomancy: Automated Performance Enhancement through Data Placement Optimization

Oceane Bel[1] • Kenneth Chang[1]
Nathan R. Tallent[3] • Dirk Duellmann[4]
Ethan L. Miller[1,2] • Faisal Nawab[1] • Darrell D. E. Long[1]
[1]University of California, Santa Cruz • [2]Pure Storage
[3]Pacific Northwest National Lab • [4]CERN

Motivation: Data layout challenges

Adjusting a system's data layout presents a variety of challenges

- **Large search space of potential data layout**
  - Many pieces of data to place between the devices
  - Any storage device can hold all the data potentially
  - Not all data layout will have similar performance
    - Compared to other data layout
    - Over different timesteps
- **Transfer overhead**
  - Not practical to shuffle the entire layout of a system's data every day
  - Transfer overheads must not exceed any predicted performance gains

Keep track of changing performance of individual components

- Data can then be routed to the best component that serves the workload

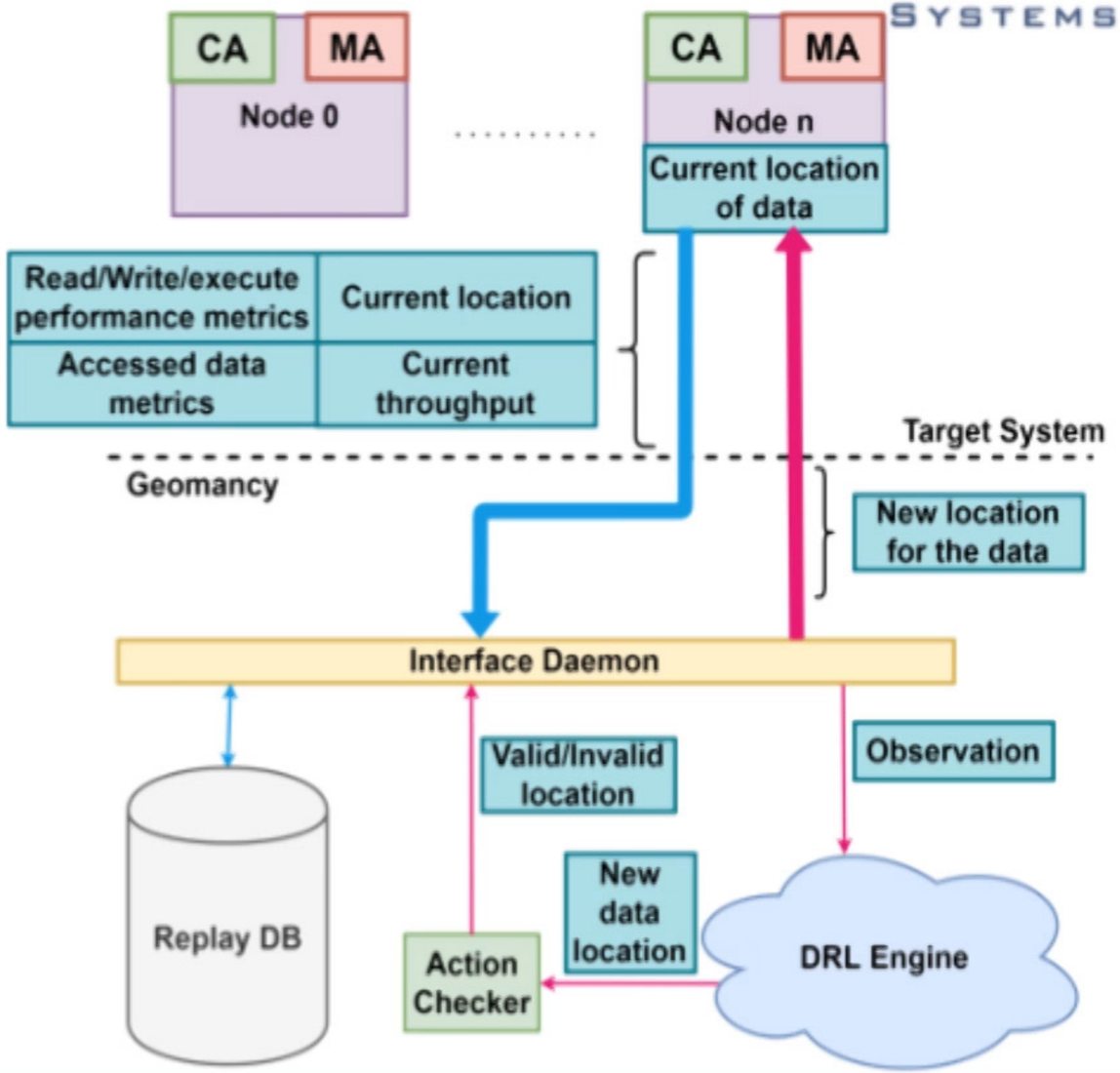
Our solution

- **Geomancy will use machine learning to**
  - Observe and learn from the executing workload on a target system
  - Use the observations to propose data layouts
    - To improve the target system's performance

- **It will learn how a target system's workload varies over time**
  - Changes in demand require different data layouts
  - Geomancy suggests the best layout for the situation

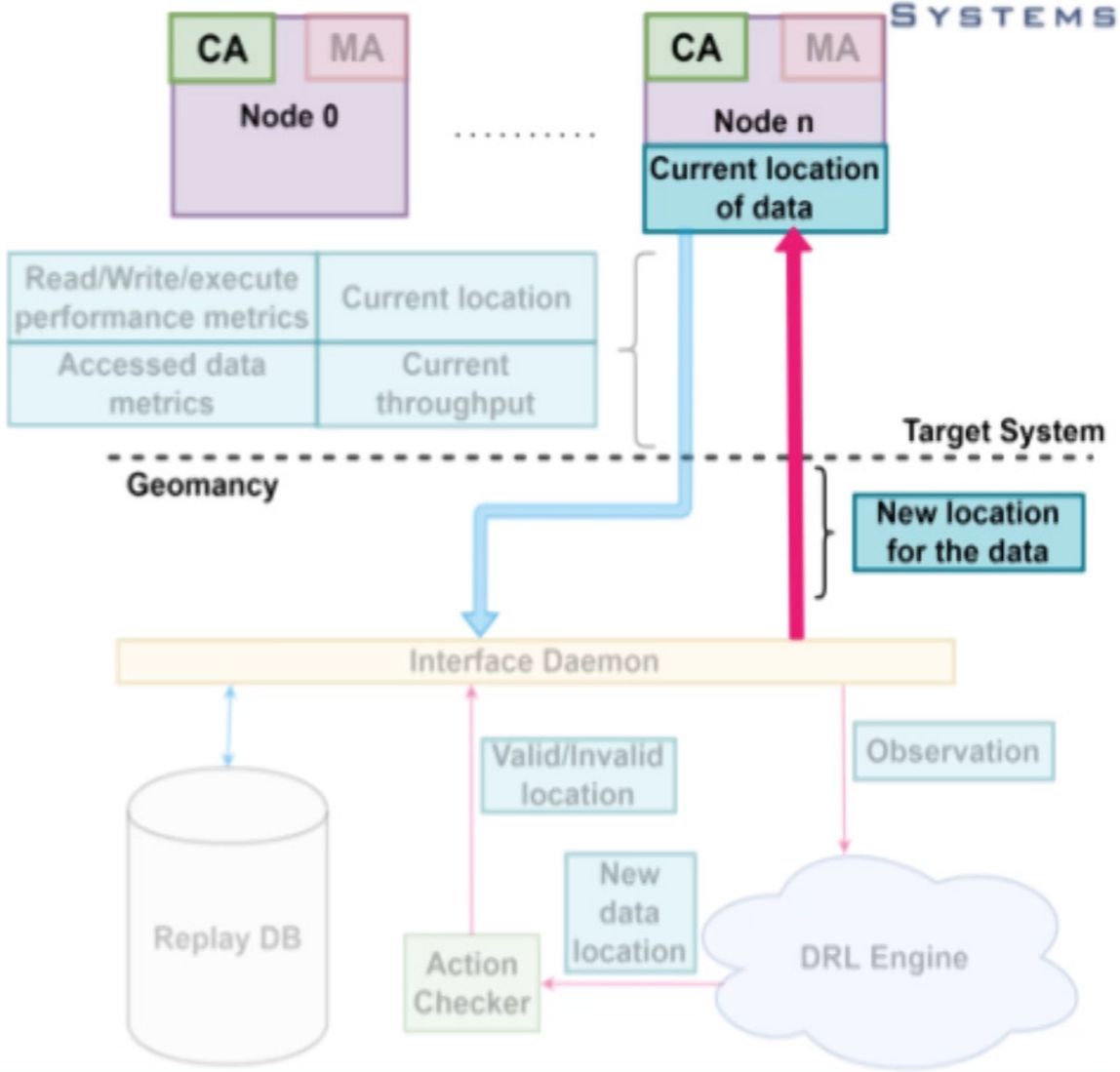
Geomancy (High level)

- **Uses file access patterns:**
  - File access patterns affect total I/O throughput of the system
  - These patterns show how the system experiences workloads over time
  - This information is used to create data movement schedules
- **Keeps track of previous and present locations of data**
  - Overhead of moving data is accounted in measured throughput values



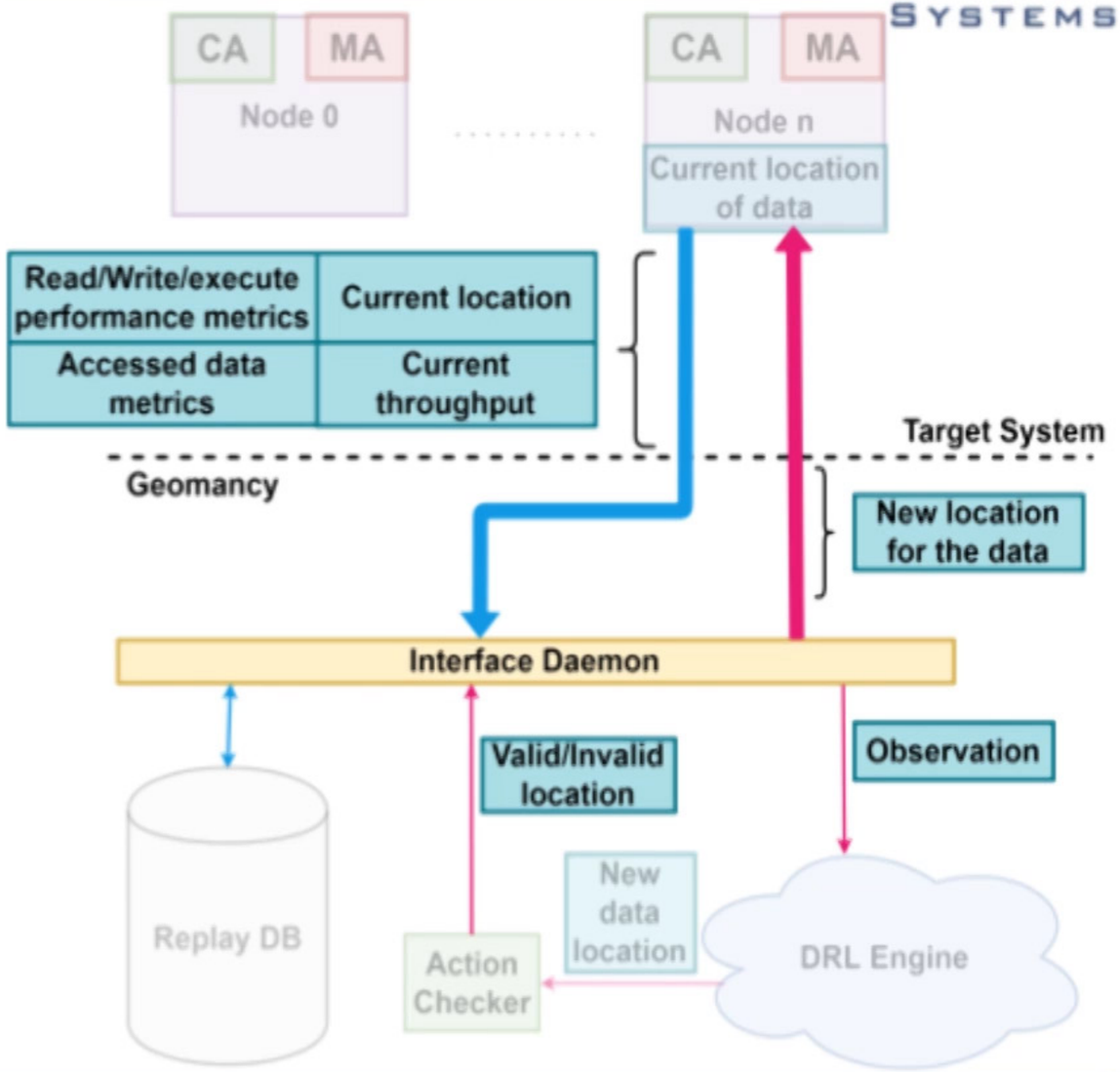
Control Agents

- **Control Agent (CA)**
  - Also located on every storage mount
  - Receives new location and ID for specific data and moves data
- **Benefits:**
  - Allows for all data to be moved in parallel
  - Prevent the system from waiting for another piece of data to be moved



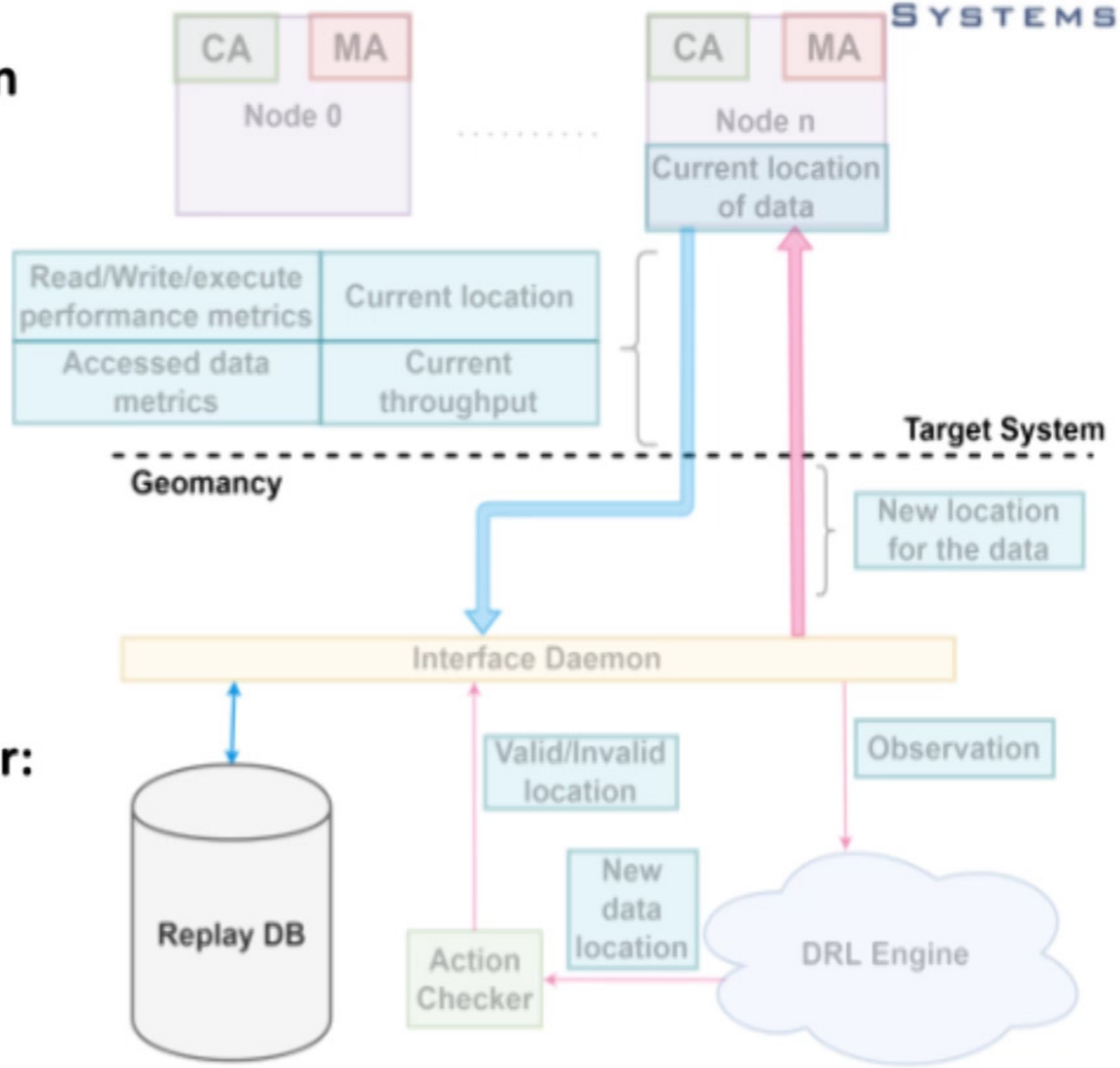
Interface Daemon

- **Networking middleware**
  - Facilitates requests between the target system and Geomancy
  - Allows for data gathering to happen at the same time as data movements
- **Gathers and processes any data needed by the DRL Engine for training/prediction purposes**
  - Allows for neural network to be trained at the same time as data is moved in the system



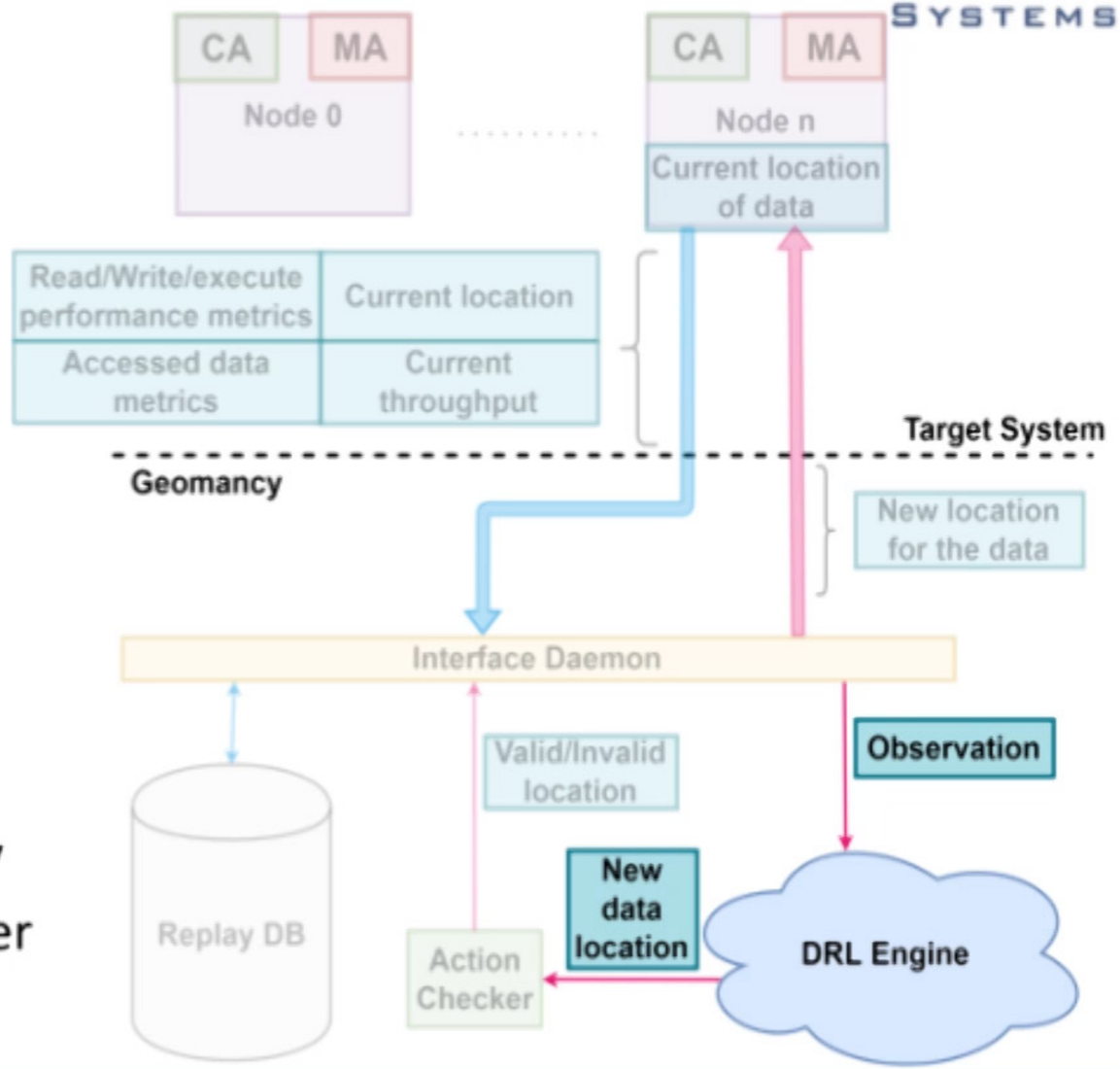
ReplayDB

- **"Replay" means getting past information for the neural network's training**
- **Stores data received from target system**
- **DRL (Deep Reinforcement Learning) Engine**
  - Gets data from the Replay DB
  - Does not interfere with the target system
- **You can put all of the replay data into memory, but a persistent DB is useful for:**
  - Tolerating system faults
  - Reproducing training results
  - Debugging during development



Deep Reinforcement Learning (DRL) Engine

- **Our current neural network is small because the PNNL data set is small, and having a larger network will quickly over-fit**
- **An input size of X features will require a layer size of (X,1), X is the number of performance features presented to the neural network**
  - More neurons will make it find incorrect trends
  - Dense layers will allow us to identify which features affect each other over time



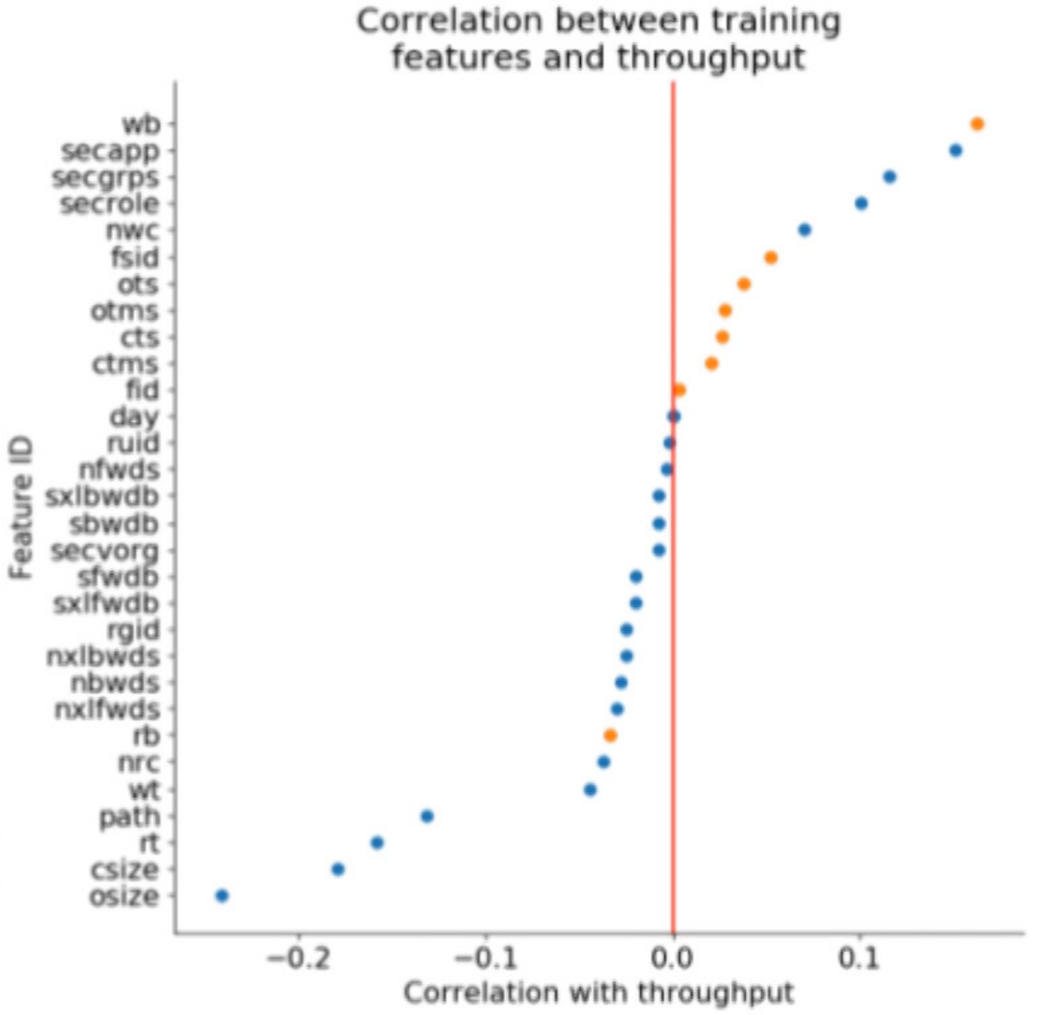
Training features selection

- **Selected features:**
  - commonly found on systems
  - non-zero correlation
- **Open timestamp (ot) and closed timestamp (ct) in seconds (s) and milliseconds (ms)**
  - Measure the length of the access
- **Read bytes (rb) and write bytes (wb)**
  - Replaced size of the data when it is opened (osize)
  - Model the overhead of moving data

Current and future location of the data (fsid)

- Model the data changing location

The data ID (fid)



Models tested

Z = number of training metrics. We used 6 training features.

Model ID	Model description
1	*16Z(Dense) ReLU, 8Z(Dense) ReLU, 4Z(Dense) ReLU, 1 (Dense) Linear*
2	(4X) 16Z(Dense) ReLU, 1(Dense) ReLU
3	(5X) 16Z(Dense) ReLU, 1 (Dense) ReLU
4	(5X) Z(Dense) ReLU, 1 (Dense) ReLU
5	*Z(SimpleRNN) ReLU, 4Z(Dense) ReLU,Z(Dense) ReLU, 1 (Dense) Linear*

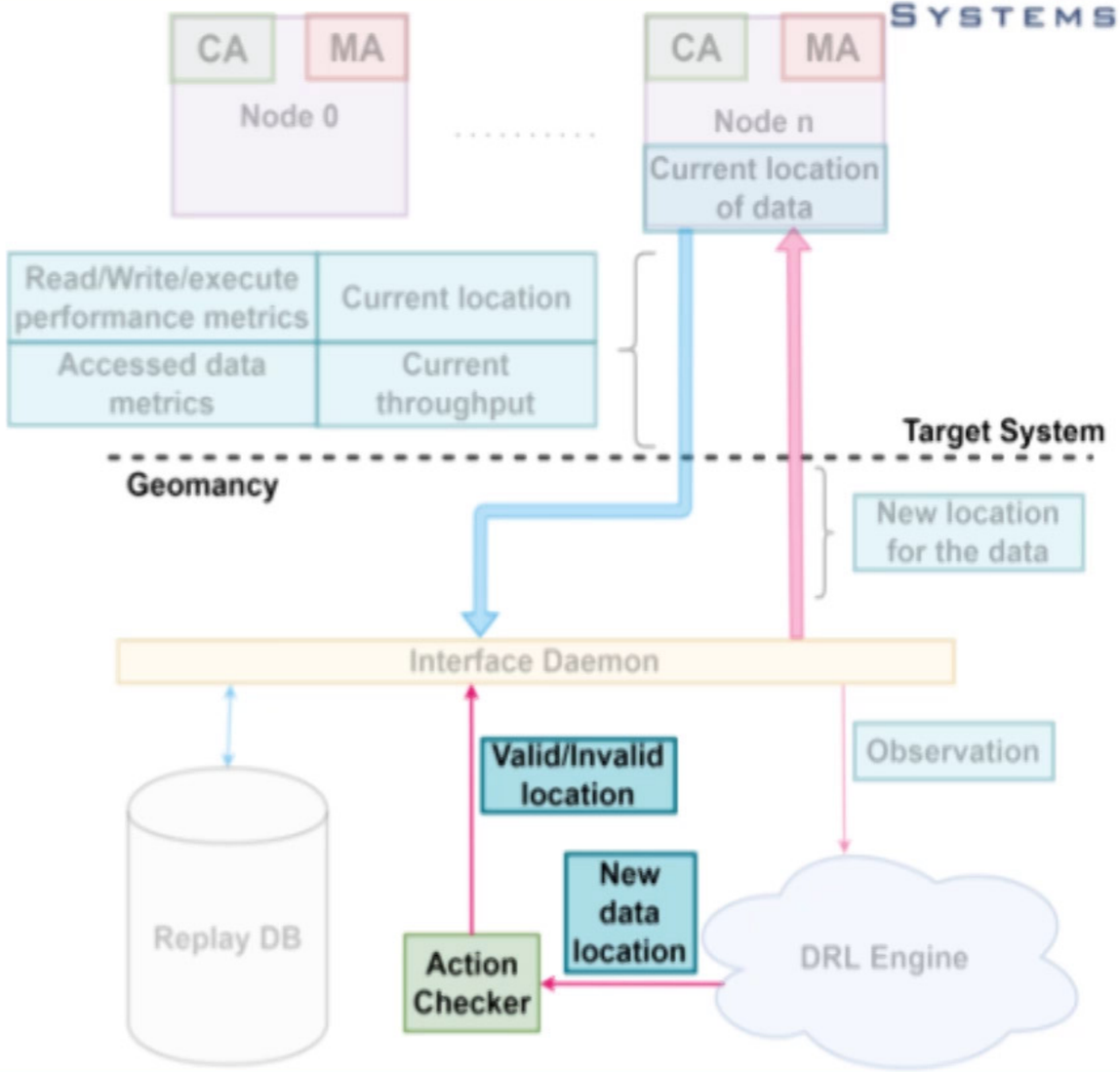
Models prediction accuracy

CENTER FOR RESEARCH IN STORAGE SYSTEMS

Model ID	Prediction mean absolute error (%)	Training time (s)	Prediction time (ms)
1	*18.88 ± 16.92*	*25.657 ± 0.801*	*55.4 ± 3.6*
2	17.63 ± 15.95	40.266 ± 0.341	70.0 ± 3.3
3	17.72 ± 16.02	47.956 ± 0.447	80.6 ± 3.7
4	18.50 ± 16.42	23.822 ± 0.498	61.0 ± 2.8
5	*18.77 ± 16.83*	*27.102 ± 0.807*	*78.0 ± 3.3*

Action checker

- Checks to see if the location that the data will be moved to is valid
- The Action Checker will look at a configuration file for a storage mount's restrictions
- Valid means
  - It exists
  - It is not failing
  - Data is allowed to exist there



Setup

- **Pacific Northwest National Laboratory (PNNL) has provided us a system to run a prototype of Geomancy**
  - Scientific simulation using 24 files on their replicated BELLE II system
  - Simulates gathering information from a particle collision and analyzing it

- **To simulate many storage device candidates, the system contains an NFS mount, a temporary RAID 1 mount, a RAID 5 mount, a Lustre file system, a SunRPC system, and an externally mounted hard disk drive**
  - RAID 5: highest read speed but takes a large penalty when writing data
  - Externally mounted HDD: slowest mount during regular workloads
  - NFS home mount: connected via 10 Gbit Ethernet

Discussion on results

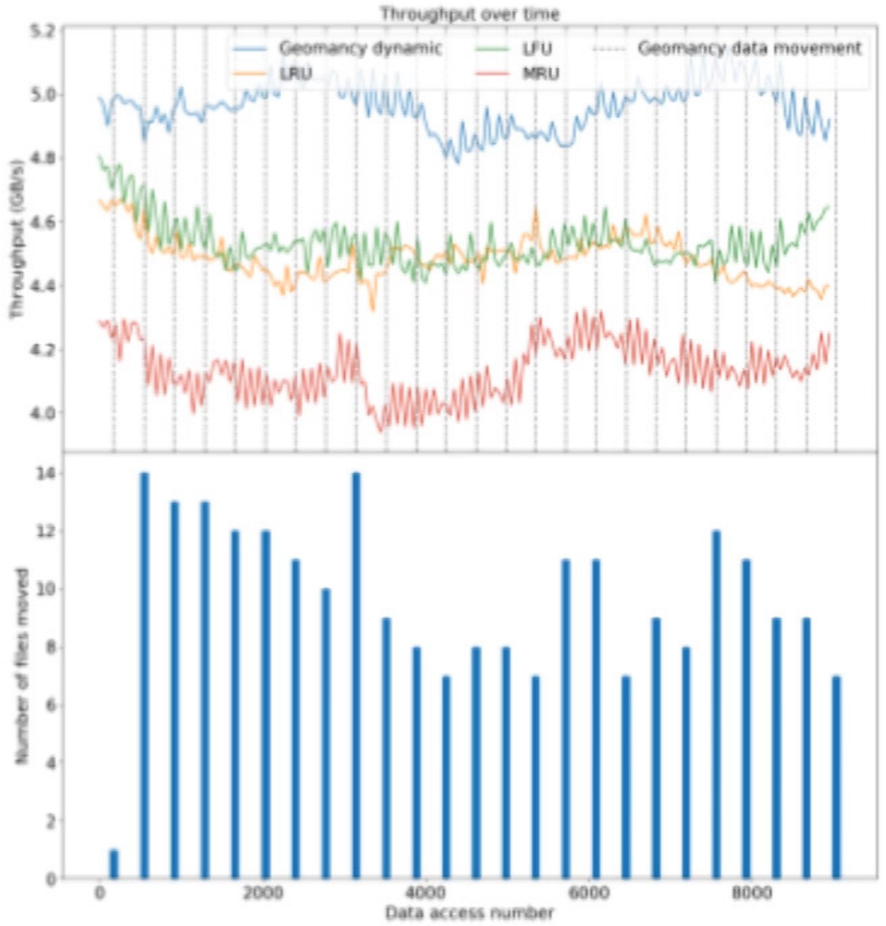
- **"Geomancy dynamic"**
  - the throughput evolution over time when Geomancy positions the data
- **"Geomancy data movement"**
  - the timestamp when geomancy moves data in the target system
- **"LFU, LRU, MRU" (Least Frequently Used, Least Recently Used, Most Recently Used)**
  - Organizes the data and spreads it across the mounts in the target system evenly
  - LFU, LRU and MRU will move the data at the beginning of each simulation
- **"Random" tests**
  - Randomly shuffle the data in the system
- **"Dynamic" tests**
  - Geomancy relocates data every 5 simulations to lower overhead
  - Other dynamic test relocate every simulations

Static test

- the data is placed before the test start and never move

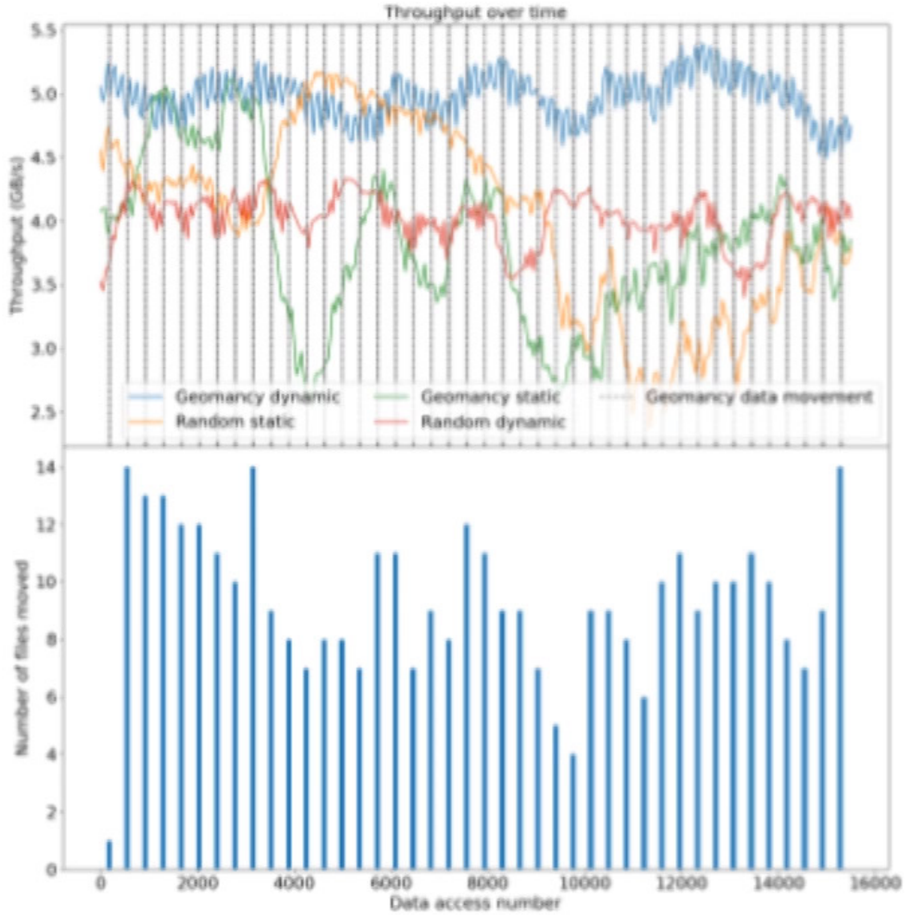
Results (1)

- Simulation has 24 files
- Data access = the time the file is opened
- Geomancy moves data every 5 simulations
- LRU, LFU and MRU move data every simulation



Results (2)

- Simulation has 24 files
- Data access = the time the file is opened
- Geomancy moves data every 5 simulations
- LRU, LFU and MRU move data every simulation



Future Work

- **Use Geomancy to move data in a multi threaded environment**
  - Commonly found in large scientific systems
- **Develop a new module which signals the action checker with the file ID that can be moved and the timestamp when it can be moved**
  - Determine a time when the file is not being heavily accessed so that it can be moved without negatively impacting the workloads on the system
- **Validate our feature selection for training our neural network**
  - Potentially identify new features which can add additional information to the performance evolution as workloads on the system change
- **Update the neural network of Geomancy to handle the new data**
  - Test out combination of RNN networks and dense networks

Acknowledgements

We would like to thank our collaborators James Byron and Daniel Bittman as well as other students from the Center for Research in Storage Systems at the University of California at Santa Cruz for their help in reviewing this paper. Additionally, we would like to thank our collaborators at Pacific Northwest National Lab and CERN for providing usage of their systems and workloads. We are grateful for funding support from the U.S. Department of Energy's (DOE) Office of Advanced Scientific Computing Research as part of "Integrated End-to-end Performance Prediction and Diagnosis." We also thank NVIDIA Corporation for their donation of a TITAN Xp GPU which will be used as part of the development of Geomancy. This research was supported by the NSF under grant IIP-1266400 and by the industrial members of the NSF IUCRC Center for Research in Storage Systems.

Thank you

CENTER FOR RESEARCH IN STORAGE SYSTEMS

Collaborators

- Oceane Bel
- Kenneth Chang
- Dirk Duellman
- Professor Ethan L. Miller
- Professor Faisal Nawab
- Professor Darrell D. E. Long