DATA DRIVEN HIGH PERFORMANCE DATA ACCESS

Presented by Dusan Ramljak, College of Engineering

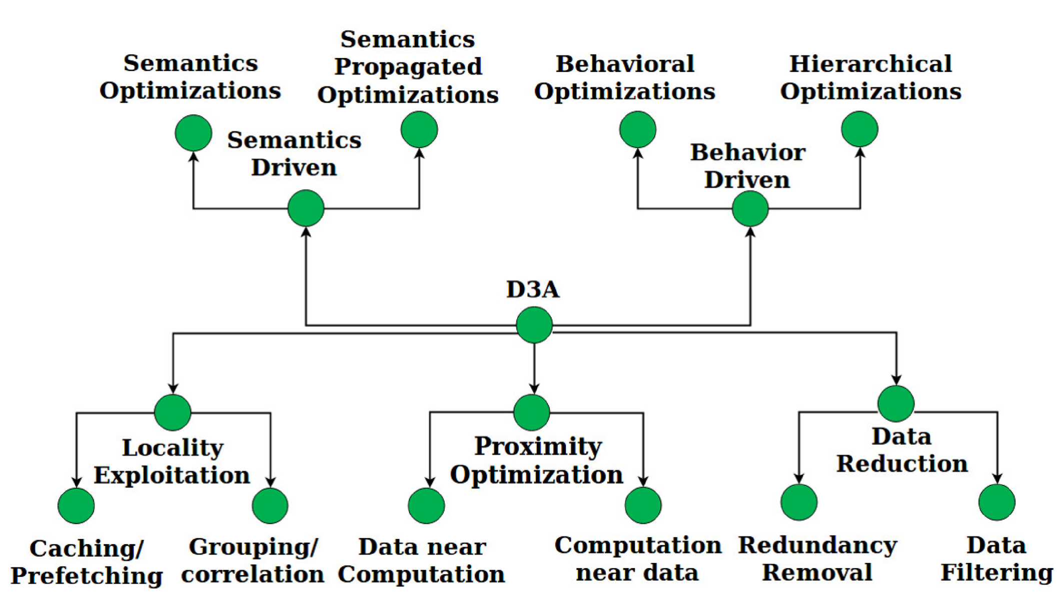
Collaborators: Jit Gupta, Anis Alazzawe, Tanaya Roy, Madhurima Ray, Amitangshu Pal, Spencer Melnick, Andrew Posmontier, Tommy Tan, Pavana Pradeep, Alexey Uversky, Jesse Friedman, Krishna Kant Temple University
Ayman Abouelwafa, Doug Voigt, Deepak Abraham tom – HPE, Tony Floeder, Jeremy Swift – Dell EMC, Valentin Kuznetsov – Huawei, Deepak Kenchammana-Hoskote - Salesforce

This research has been conducted as a part of NSF CRIS IUCRC, and has been supported by HPE, Dell EMC, Huawei and Salesforce

Motivation

- Data size already enormous and growing:
  – Over 90% of world data generated in last 2 years
- Large scale systems highly distributed
  – Resource sharing brings opportunities and challenges
- Strain and opportunity for data centers and cloud storage providers
  – Efficient pattern discovery and description
  – Observed predictability of complex patterns
- Intelligent management of storage tiers is crucial

Data Driven high performance Data Access (D3A) overview



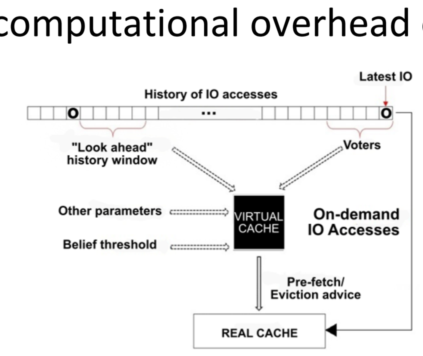
Locality Exploitation Foundations

- Belief – estimate of a probability of a hit
- Conditional probability x being in window of calculated:

$$P(\{x\}|E = o) = \frac{c_{o-x}}{c_{xp}} * \frac{f_x}{f_o}$$

- {x}: x is in the **"look-ahead"** window
- {x}|E = o: {x} conditioned by current object o
- c_{o-x}: count of x in "look-ahead" windows of o
- c_{x_p}: count of x in all "look-ahead" windows
- f_o, f_x : frequency of o, and x respectively

- Modular approach allows for
  – Decoupling decision making from cache management
  – Reducing the computational overhead of belief computation

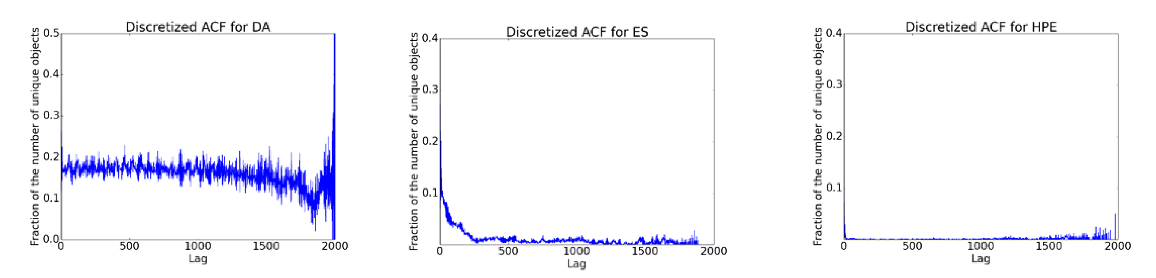


- Framework exploits
  – Various levels of successor predictability
  – Indirectly exploits access history usage patterns
  – Local/global view combination expected to provide best decisions

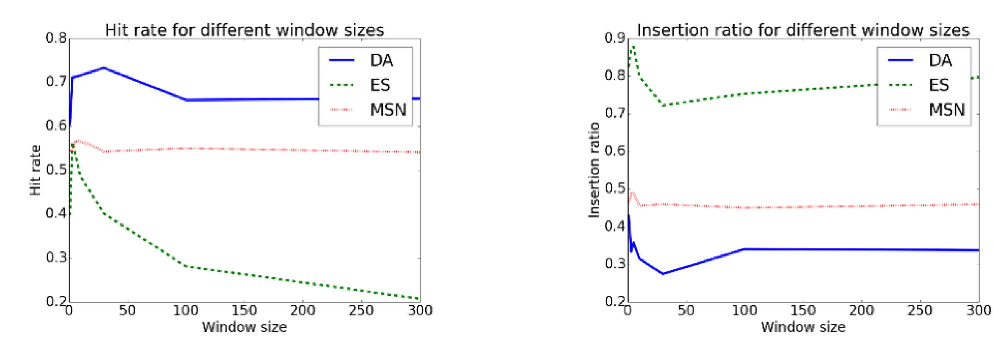
- Obtaining good results largely depends on parameters
  – For the perturbations of the same workload able to adapt
  – For the drastic change in workload/different workload - retrain

Locality Exploitation Evaluation

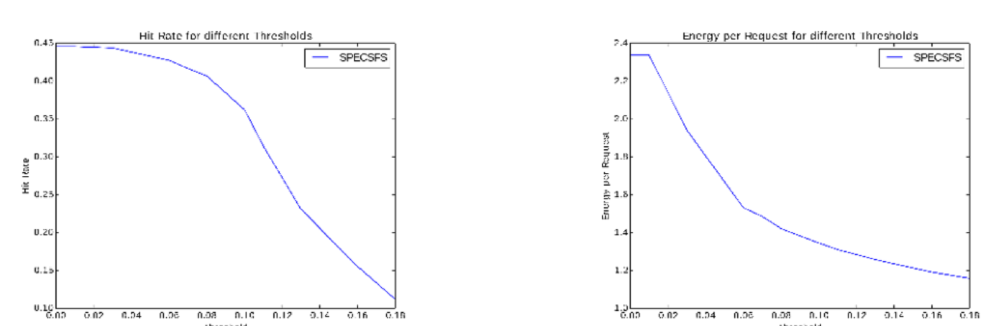
- Microsoft block traces (available at SNIA website)
  – Primarily Disk IO, FileIO traces with following unique characteristics



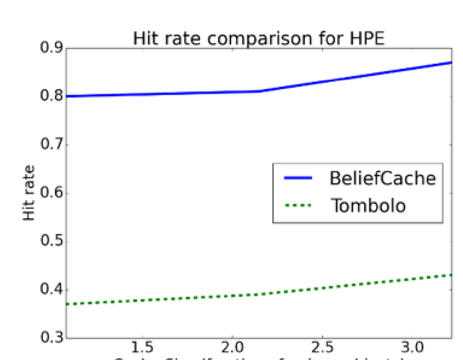
- SPEC SFS suite
  – Video Data Acquisition workload (HPE above)
  – Virtual Desktop Infrastructure workload
- Noticeable trends
  – For workloads exhibiting short and medium term correlations



  – For workloads exhibiting long term correlations

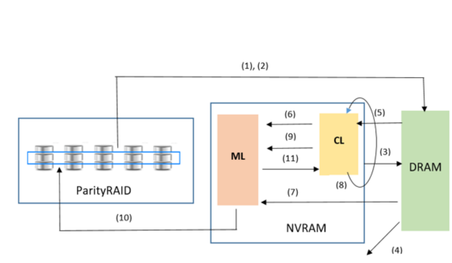


- Complex non-sequential patterns captured better than state-of-the-art framework for optimizing cloud storage gateways

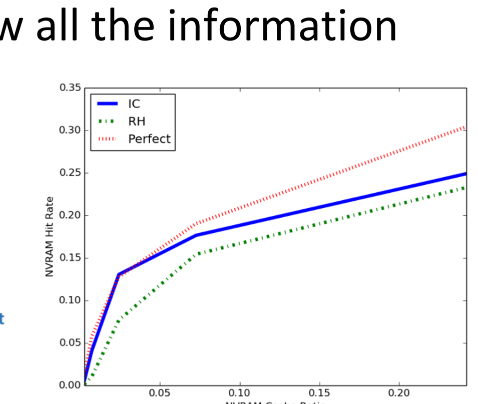


- Further improvements are possible by
  – Understanding the effects of phase change on caching characteristics
  – Using the phase change detection mechanism to automate retraining

- Optimized stripe cache splitting between DRAM and NVRAM
- Balance trade-off between Performance, Resilience, Endurance
- Built efficient, flexible mechanism that adapts to
  – Amount and speed of DRAM and NVRAM
  – Workload Characteristics (VDI trace on picture below)
- IC – instance of locality exploitation evaluation
- RH – instance of currently applied policies
- Perfect – what we could get if we knew all the information



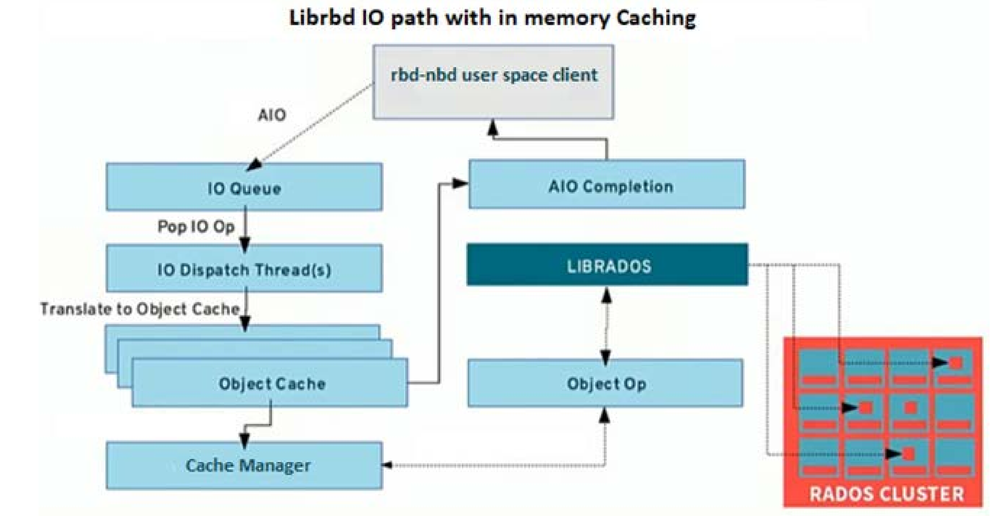
(1) Prefetch
(2) Demand Fetch
(3) NVRAM Fetch
(4) DRAM Clean Eviction
(5) DRAM Clean Eviction Store
(6) NVRAM CL Persist
(7) DRAM Persist
(8) NVRAM Clean Eviction Store
(9) NVRAM Clean Eviction Persist
(10) NVRAM Writeback
(11) NVRAM Writeback Store



Locality Exploitation Implementation

- Previous prefetching
  – Read ahead of extra bytes from same object upon sequential access
- Challenges
  – Keeping track of object id is not enough
  – Requires asynchronous requests
  – No fixed unit of request
  – Entire object may be to big to cache
  – Complex striping

Librbd IO path with in memory Caching



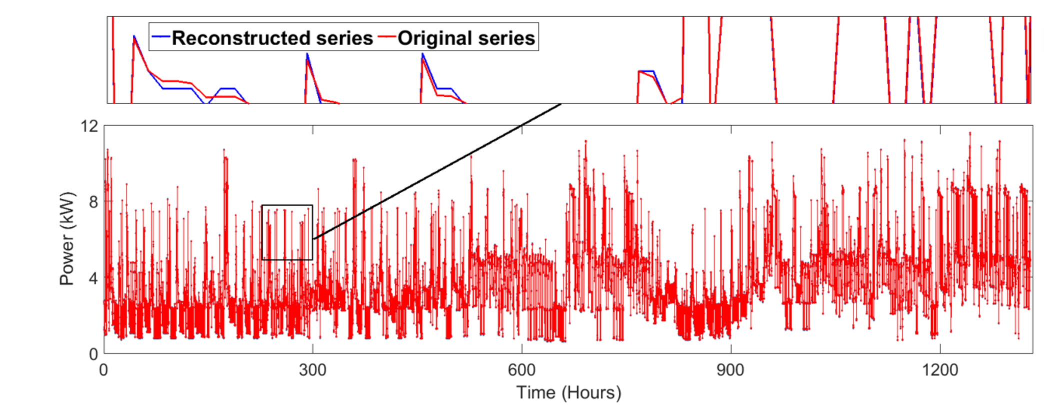
Data Reduction Foundations

- The increasing proliferation of the IoT devices and systems
- Large amounts of highly heterogeneous data to be collected
  – Multiple sensors "readings" at successive time points
  – A continuous stream of data is being generated
  – With n sensors each sample will be an n-element vector
  – Use cases when it is acceptable to scale the received values
    • Discretize the value of the i-th sensor into bi bins, for a suitable b
- Provenance information is required
  – Describe the data in a sufficiently intuitive way - easily usable
- Lossy data compression mechanisms necessary
  – Provide a trade-off between
    • Compression ratio and
    • Data usefulness for offline statistical analysis
  – Preserve the information about the sequence of events
    • Crucial in IoT context

- Ensure that the sequence of events can be derived (or inferred) from the compressed data
- Keep some events intact and approximate other events to improve the compressibility
- Two stages of the method
  – Use the SQS1 to generate a summarization of the data S
  – Approximately match the remaining events with the patterns in S
    • A matching threshold τ is the Euclidean distance between the subsequence and the existing pattern from S
- We consider the following two ways to apply SQS:
  – (Correlated AVSC-C) Directly compress each vector
    • Retains the correlation across streams perfectly,
    • It may show poor tradeoff of compressibility and accuracy
  – (Uncorrelated AVSC-NC) Compress i-th stream separately
    • Different approximations across vectors might lead to approximate information about corresponding streams
    • Expected to yield better compression at the cost of confounding the correlation across the attributes

AVSC Discussion

- Lossy compression shows gains over the plain lossless compression
  – For a specified amount of accuracy for purposes of identifying the state of the system
- Trade-off in between the compressibility and fidelity
  – With respect to distance threshold τ
    • Fidelity increases with small thresholds τ at the cost of poor compressibility
- Observed behavior is domain specific
  – There is a strong correlation between weather and energy usage
  – Discretization smoothed the differences between the vectors



Conclusions and Future Work

- Opportunities and challenges in developing data drive high performance data access methods
  – Leverage data properties and relationships between data items
    • Minimize data movement, latency, energy consumption
- Rapidly emerging data intensive applications should focus on
  – Proximity optimizations leading to reduction of distance between data and computations
  – Data reduction by removing redundancy in data
  – Using sparse representations of data
  – Impact of data access mechanisms on
    • Resilience
    • Energy consumption
    • Storage usage
    • Enablement of new classes of data driven applications

References

- Ramljak, D. "Data Driven High Performance Data Access", PhD Thesis, 2019
- Ramljak, D., Tom, D., Voigt, D., Kant, K., "Modular framework for Data Prefetching and Replacement at the Edge", In International Conference on Edge Computing, Research track of SCF, Seattle, USA June 25 - June 30, 2018
- Ramljak, D., Kant, K. "Belief-Based Storage Systems," In The HotStorage '17, WACI session, Santa Clara, July 10-11, 2017
- Ramljak, D., Alazzawe, A., Uversky, A., Kant, K. "Belief-Based Data Prefetching and Replacement in Storage Systems," In The 15th USENIX Conference on File and Storage Technologies (FAST), Work in progress (WiP) session, Santa Clara, Feb 27 - Mar 2, 2017
- Ramljak, D., Davey, A., Uversky, A., Roychoudhury, S., Obradovic, Z. (2015) "Casting a Wider Net: Data Driven Discovery of Proxies for Target Diagnoses," AMIA 2015 Annual symposium, San Francisco, Nov. 14 - 18 2015
- Ramljak, D., Davey, A., Uversky, A., Roychoudhury, S., Obradovic, Z. (2015) "Hospital Corners and Wrapping Patients in Markov Blankets," 4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining, Vancouver, Canada, April 30 - May 02, 2015
- Zhang, Y., Ramljak, D., Luo, X. "Using the Entire Network of Big Data to Predict Individual Behavior", American Marketing Association 2014 summer academic conference
- Uversky, A., Ramljak, D., Radosavljevic, V., Ristovski, K., Obradovic, Z. (2014) "Panning for Gold - Using Variograms to Select Useful Connections in a Temporal Multigraph Setting," Social Network Analysis and Mining, 4:211, July 2014
- Uversky, A., Ramljak, D., Radosavljevic, V., Ristovski, K., Obradovic, Z. (2013) "Which Links Should I Use? A Variogram Based Selection of Relationship Measures for Prediction of Node Attributes in Temporal Multigraphs," Proc. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, Canada, Aug. 2013
- Ghalwash, M., Ramljak, D., Obradovic, Z. (2015) "Patient-Specific Early Classification of Multivariate Observations," International Journal of Data Mining and Bioinformatics, Vol. 11, No. 4, 2015
- Ghalwash, M., Ramljak, D., Obradovic, Z. (2012) " Early Classification of Multivariate Time Series Using a Hybrid HMM/SVM model," Proc. 2012 IEEE International Conference on Bioinformatics and Biomedicine, Philadelphia, PA, Oct. 2012
- Crainic, T. G., Davidović, T., Ramljak, D., (2014) "Designing parallel meta-heuristic methods" , High Performance and Cloud Computing in Science and Education, Despotović-Zrakić, M., Milutinović, V., Belić, A., (eds.), IGI-Global, pp. 260-280, 2014. CIRRELT Research Report CIRRELT-2012-28
- Davidović, T., Šelmić, M., Teodorović, D., Ramljak, D., (2012) "Bee Colony Optimization for Scheduling Independent Tasks to Identical Processors," Journal of Heuristics, Volume 18, Issue 4, pp 549-569
- Davidović, T., Ramljak, D., Šelmić, M., Teodorović, D., (2011) "Bee Colony Optimization for the p-Center Problem," Computers and Operations Research, 38(10) pp. 1367--1376
- Davidović, T., Ramljak, D.,Šelmić, M., Teodorović, D., (2011) "MPI Parallelization of Bee Colony Optimization," Proc. 1st Int. Symp and 10th Balkan Conf. on Operational Research, BALCOR 2011, Vol. 2, pp. 193--200, Thessaloniki, Greece, Sept. 22-24
- Davidović, T., Ramljak, D., Šelmić, M., Teodorović, D., (2010) "Parallel Bee Colony Optimization for Scheduling Independent Tasks on Identical Machines," Proc. 37th Symp. on Operational Research, SYM-OP-IS 2010, pp. 389--392, Tara, Sept. 21-24
- Teodorović, D., Davidović, T., Šelmić, M., Ramljak, D., (2010) "An Application of a Meta-heuristic Algorithm to p-center Location Problem " (in Serbian), Proc. Symp. on information technology, YUINFO 2010, (on CD 026.pdf), Kopaonik, March 03-06