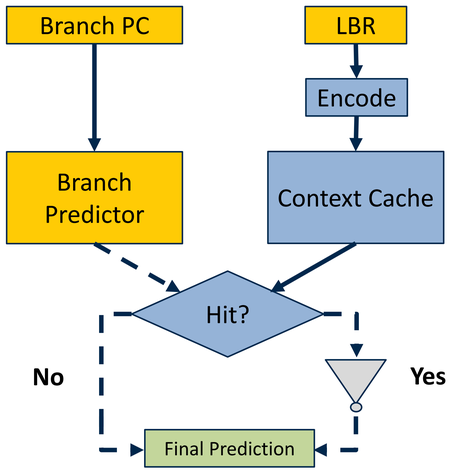
Understanding Branch Prediction in Datacenter Applications

PRESENTER:

Muhammed Ugur

- Modern datacenter applications consist of large instruction footprints, increasingly complex logic, and deep software stacks [1]
- This complexity makes it more difficult for online branch predictors to learn intricate patterns and correlations in branch history
- We propose the use of offline profiling and hardware/software co-design to tackle these growing challenges in branch prediction

PROPOSED MECHANISM



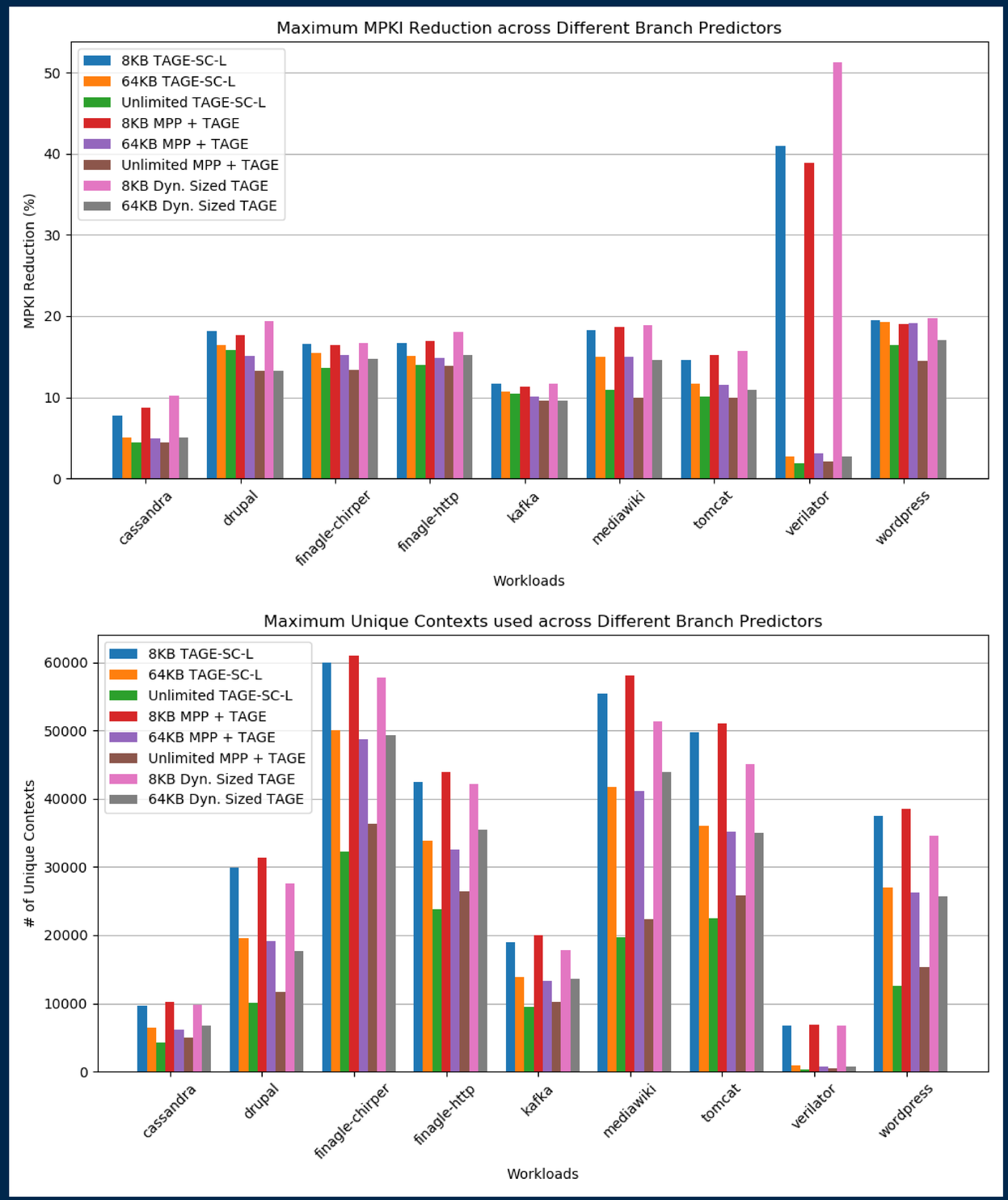
- The Context Cache stores encoded global history patterns that were deemed hard-to-predict from application profiles
- Contexts are encoded by concatenating the two least significant bits of the last 32 branches in the Last Branch Record

DATACENTER WORKLOADS

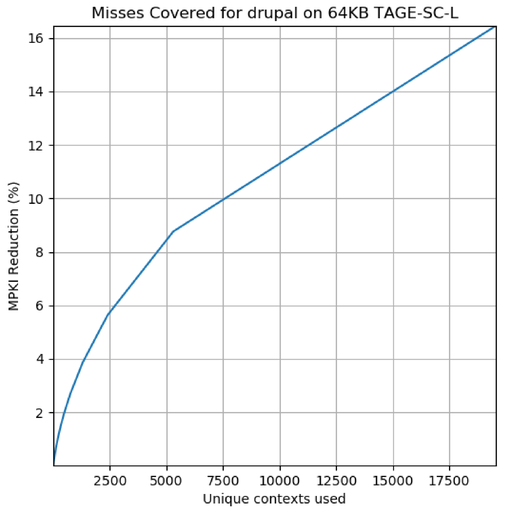
Workload	Description
Cassandra	NoSQL DBMS used by Netflix/Uber
Drupal	PHP-based CMS (Facebook's OSS-performance)
Finagle-Chirper	Micro-blogging service (Java Renaissance)
Finagle-HTTP	HTTP server (Java Renaissance)
Kafka	Apache's stream-processing platform used by Uber/LinkedIn/Airbnb
Mediawiki	Open-source Wiki engine (Facebook's OSS-performance)
Tomcat	Apache's open-source Java web server
Verilator	Tool for simulating custom hardware designs used by Intel/ARM
Wordpress	PHP-based CMS (Facebook's OSS-performance)

[1]

Prefetching branch metadata through hardware/software co-design and profiling significantly reduces MPKI



CONSIDERING OVERHEAD



- ~50% of the ideal MPKI reduction can be gained by targeting the most impactful contexts, reducing storage overhead
- Re-evaluating the size of each context (64 bits) can also significantly reduce storage

ADDITIONAL EVALUATION

- Evaluating our datacenter traces on the software-only Big-BranchNet [2] provided 10.41% avg. MPKI reduction compared to our design's 12.36% avg. on 64KB TAGE-SC-L

CBP 2017 Trace	MPKI Reduction	Unique Contexts
SHORT_MOBILE-44	32.46%	1799
SHORT_SERVER-139	12.59%	11876
SHORT_SERVER-6	12.27%	12527
SHORT_MOBILE-43	9.76%	466
LONG_SERVER-4	8.72%	15718
SHORT_MOBILE-40	8.25%	176
SHORT_MOBILE-23	7.55%	26216
SHORT_MOBILE-12	7.37%	55
LONG_MOBILE-12	7.33%	22520

Muhammed Ugur
Tanvir Ahmed Khan
Krishnendra Nathella
Dam Sunwoo
Daniel A. Jiménez
Baris Kasikci

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

[1] Khan, T. A., Brown, N., Sriraman, A., Soundararajan, N. K., Kumar, R., Devietti, J., Subramoney, S., Pokam, G. A., Litz, H., & Kasikci, B. (2021). Twig: Profile-Guided BTB Prefetching for Data Center Applications. *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 816–829.

[2] Zangeneh, S., Pruett, S., Lym, S., & Patt, Y. N. (2020). BranchNet: A Convolutional Neural Network to Predict Hard-To-Predict Branches. *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 118–130.