A Two Level Neural Approach Combining Off-Chip Prediction with Adaptive Prefetch Filtering

Alexandre Valentin Jamet[1], Georgios Vavouliotis[2], Daniel A. Jiménez[3], Lluc Alvarez[1], and Marc Casas[1]

[1]*Barcelona Supercomputing Center* [2]*Huawei Zurich Research Center* [3]*Texas A&M University*

1. Introduction

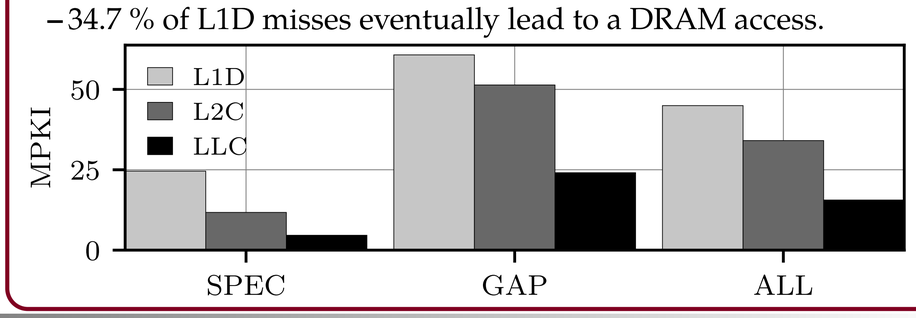
- We present TLP, a scheme combining Off-Chip Prediction and L1D Prefetch Filtering leveraging two perception predictors.
- TLP performs accurate Off-Chip Prediction and L1D Prefetch Filtering leveraging features (physical/virtual addresses, etc.).
- TLP is the first proposal targeting Off-Chip Prediction and Prefetch Filtering using an hardware **Multi-Level** Perceptron.

2. Background

- Workloads load caches, approx. 80% latency from DRAM.
- Off-Chip Prediction (Hermes[1] and LP[2]) addresses high-latency load requests.
- Prefetch Filtering (PPF) balances miss coverage and accuracy.

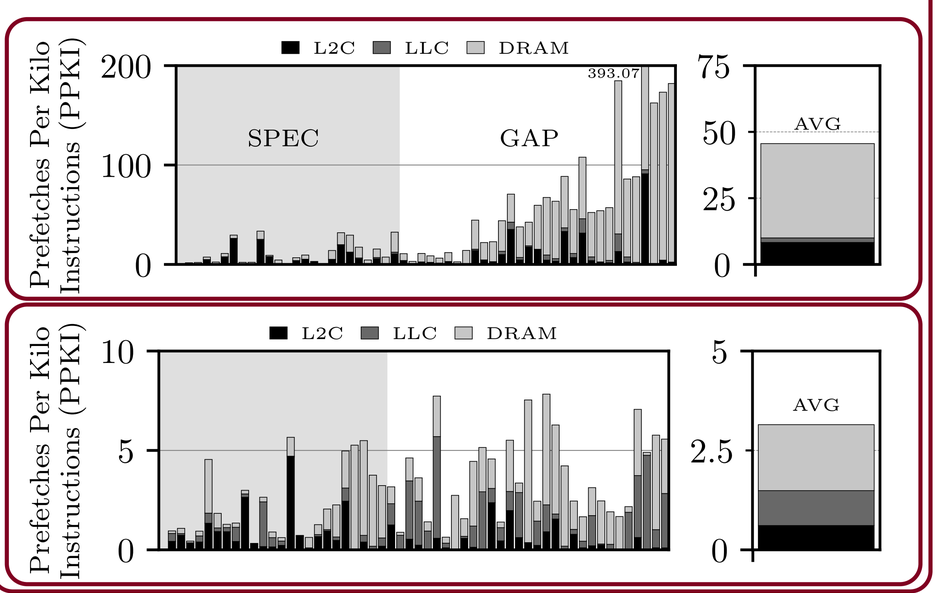
3. Cache Behavior of Modern Workloads

- Measuring MPKI in the L1D, L2C, and LLC.
  - These caches have MPKIs of 45.0, 34.1, and 15.6, respectively.
- A large portion of large working set load requests miss caches.
  - 34.7 % of L1D misses eventually lead to a DRAM access.



5. Off-Chip Prediction for L1D Prefetch Filtering

- Of the inaccurate prefetch requests, L2C serves 18.2%, LLC serves 3.8%, and DRAM serves 78%
- Off-chip prediction can aid creating an effective L1D prefetch filter.



7. Methodology

- ChampSim simulator modeling a Cascade Lake microarchitecture.
- The baseline uses prefetchers (L1D: IPCP[3]/Berti[4], L2C: SPP).
- Workloads: 24 SPEC CPU and 31 GAP (LLC MPKI > 1).

Alternative Techniques

- Hermes
- PPF
- Hermes+PPF

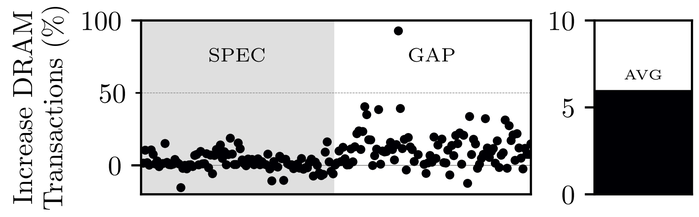
[1] Bera et al., "Hermes: Accelerating Long-Latency Load Requests via Perceptron-Based Off-Chip Load Prediction." MICRO'22
[2] M. Jalili and M. Erez, "Reducing Load Latency with Cache Level Prediction," HPCA'22
[3] S. Pakalapati and B. Panda, "Bouquet of Instruction Pointers: Instruction Pointer Classifier-based Spatial Hardware Prefetching," ISCA'20
[4] A. Navarro-Torres et al., "Berti: an Accurate Local-Delta Data Prefetcher," MICRO'22.

This work has been partially supported by the European HiPEAC Network of Excellence, by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 (contracts PID2019-107255GB-C21 and PID2019-105660RB-C22) and by the Generalitat de Catalunya (contract 2021-SGR-00763). This work is supported by the National Science Foundation through grant CCF-1912617 and generous gifts from Intel. Marc Casas has been partially supported by the Grant RYC-2017-23269 funded by MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future. Els autors agraeixen el suport del Departament de Recerca i Universitats de la Generalitat de Catalunya al Grup de Recerca "Performance understanding, analysis, and simulation/emulation of novel architectures" (Codi: 2021 SGR 00865). This research has received funding from the European High Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 (European Processor Initiative) and Specific Grant Agreement No 101036168 (EPI SGA2). The JU receives support from the European Union's Horizon 2020 research and innovation programme and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland. The EPI-SGA2 project, PCI2022-132935 is also co-funded by MCIN/AEI /10.13039/501100011033 and by the UE NextGenerationEU/PRTR.

4. Impact of Hermes

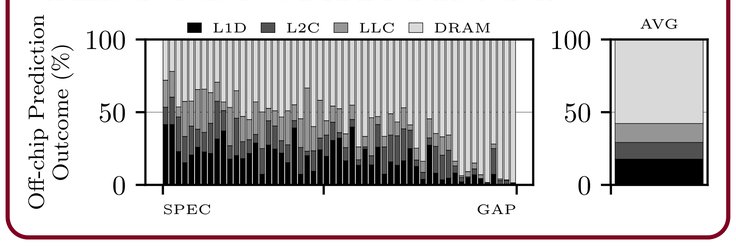
3.A. DRAM Transactions

- Within a multi-core scenario, Hermes increases DRAM transactions by 6.0%.
- Hermes significantly increases the number of DRAM accesses, especially for graph workloads.



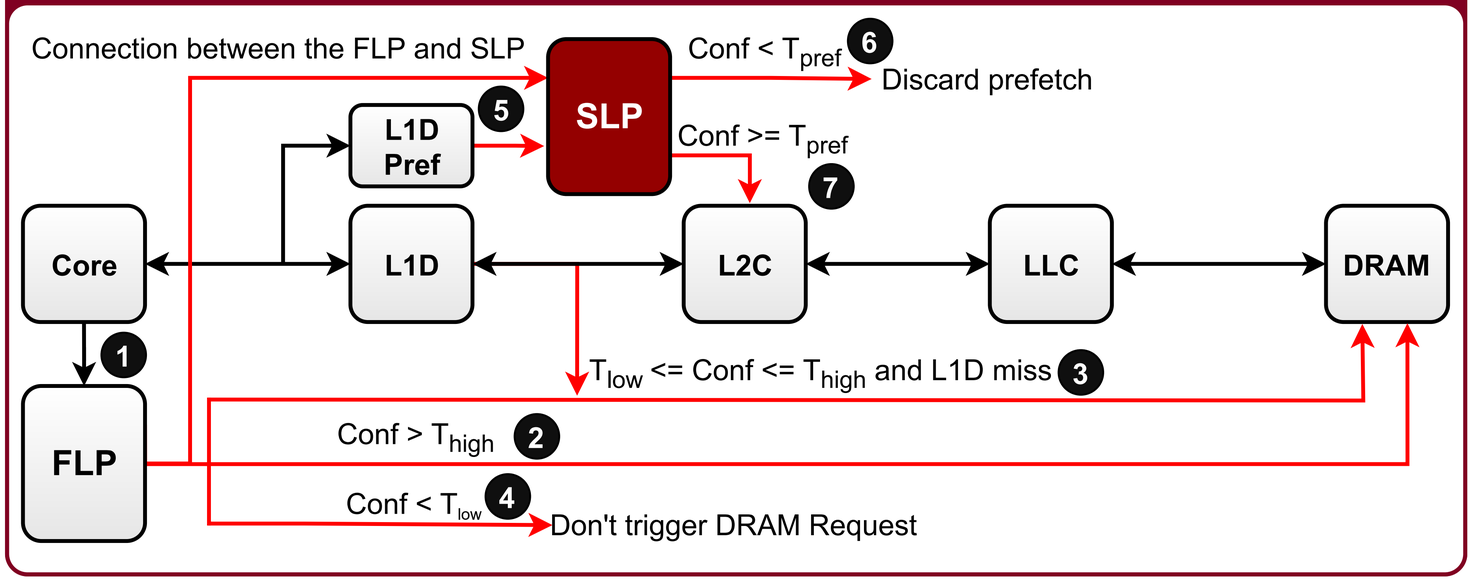
3.B. Analysis of Hermes Predictions

- 42.2% of Hermes' off-chip predictions are wrong.
- Selectively delaying Hermes off-chip predictions until the L1D lookup is resolved can reduce the number of useless DRAM transactions.



6. TLP: A Solution to Address Useless DRAM Transactions

5.A. The *Two Level Perceptron* (*TLP*) Predictor



5.C. Features and Hardware Budget

- TLP provides a Layered Perceptron design, as it connects the FLP and the SLP.
- FLP and SLP rely on the following features to build predictions:
  - PC ⊕ cacheline offset
  - PC ⊕ byte offset
  - PC + first access
  - Cacheline offset + first access
  - Last-4 load PCs
  - FLP prediction + cacheline offset
- In total, TLP requires 6.98 KB of storage.
  - FLP requires 3.21 KB of storage for its prediction tables, SLP requires 3.28 KB for its prediction tables as SLP leverages an additional feature connecting it to the FLP.
  - 0.48 KB to store additional metadata in the LQ entries and in the L1D MSHR entries.

8. Evaluation

Evaluation outcome

- In the multi-core context, using the IPCP prefetcher, **TLP** outperforms **PPF**, **Hermes**, and **Hermes+PPF**. **TLP**: 11.5%, **PPF**: -3.3%, **Hermes**: 3.0%, **Hermes+PPF**: -0.5%
- TLP significantly reduces DRAM transactions over the baseline while other techniques increase it. **TLP**: -17.7%, **PPF**: 6.5%, **Hermes**: 6.0%, **Hermes+PPF**: 13.4%
- Over a range of realistic DRAM bandwidth allocation (1.6 to 6.4 GB/s per core), TLP outperforms PPF, Hermes, and Hermes+PPF. Even in unrealistic scenarios, TLP outperforms Hermes and PPF.
- Similarly, in the single-core context, TLP outperforms all the alternative techniques.
- Finally, when evaluating TLP using the Berti prefetcher, TLP outperformrs the alternative techniques.