Advanced Data Management for Data Analysis

Stefan Manegold

Data Management @ LIACS

Group leader Database Architectures
Centrum Wiskunde & Informatica (CWI)
Amsterdam

s.manegold@liacs.leidenuniv.nl
http://www.cwi.nl/~manegold/

ADM: Agenda

- 07.09.2022: Lecture 1: **Introduction**
- 14.09.2022: Lecture 2: **SQL Recap**

  (plus Assignment 1 [in groups; 3 weeks]: TPC-H benchmark)
- 21.09.2022: Lecture 3: **Column-Oriented Database Systems (1/6) - Motivation & Basic Concepts**
- 28.09.2022: Lecture 4: **Column-Oriented Database Systems (2a/6) - Selected Execution Techniques (1/2)**

  (plus Assignment 2 [in groups; (1+)4 weeks]: Compression techniques)
- 05.10.2022: Lecture 5: **Column-Oriented Database Systems (2b/6) - Selected Execution Techniques (2/2)**
- 12.10.2022: Lecture 6: **Column-Oriented Database Systems (3/6) - Cache Conscious Joins**
- 19.10.2022: Lecture 7: **Column-Oriented Database Systems (4/6) - "Vectorized Execution"**
- ~~26.10.2022~~: ***No lecture!***
- 02.11.2022: Lecture 8: **Branch Misprediction & Predication**

  (plus Assignment 3 [individual; 2 weeks]: Predication)
- 09.11.2022: Lecture 9: **DuckDB: An embedded database for data science (1/2) (guest lecture & *hands-on*)**

  (plus Assignment 4 [individual; (1+)2 weeks]: Analysing NYC Cab dataset with DuckDB)
- 16.11.2022: Lecture 10: **DuckDB: An embedded database for data science (2/2) (guest lecture & *hands-on*)**
- 23.11.2022: Lecture 11: **Column-Oriented Database Systems (5/6) - Adaptive Indexing**
- 30.11.2022: Lecture 12: **Column-Oriented Database Systems (6/6) - Progressive Indexing**

ADM: Literature

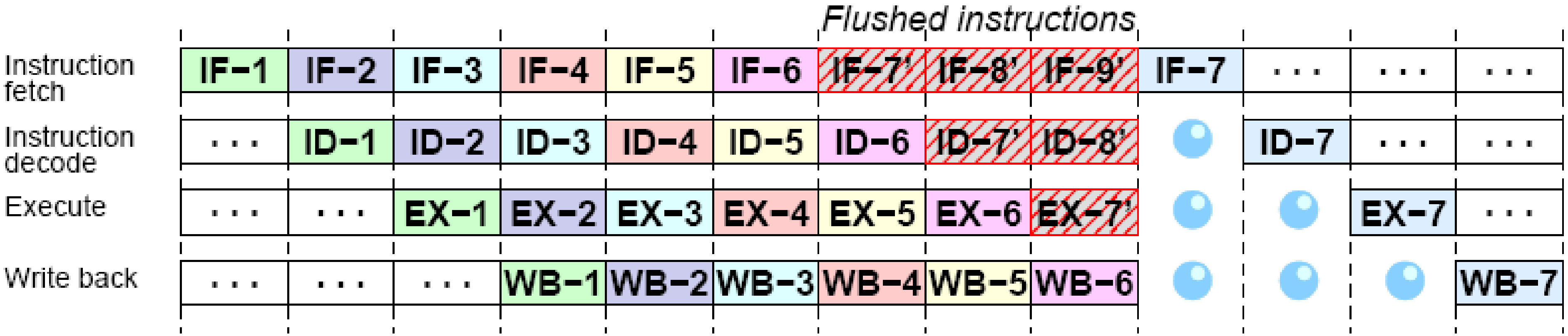
- **Branch Misprediction & Predication**
  - “Conjunctive Selection Conditions in Main Memory”. Kenneth A. Ross. PODS 2002.
  - “Selection conditions in main memory”. Kenneth A. Ross. ACM Transactions On Database Systems 29: 132-161, 2004.

Hazards

- Data hazards
  - Dependencies between instructions
  - L1 data cache misses

- Control Hazards
  - Branch mispredictions
  - Computed branches (late binding)
  - L1 instruction cache misses

Result: bubbles in the pipeline



Flushed instructions

Instruction fetch	IF−1	IF−2	IF−3	IF−4	IF−5	IF−6	IF−7'	IF−8'	IF−9'	IF−7	. . .	. . .	. . .
Instruction decode	. . .	ID−1	ID−2	ID−3	ID−4	ID−5	ID−6	ID−7'	ID−8'	●	ID−7	. . .	. . .
Execute	. . .	. . .	EX−1	EX−2	EX−3	EX−4	EX−5	EX−6	EX−7'	●	●	EX−7	. . .
Write back	. . .	. . .	. . .	WB−1	WB−2	WB−3	WB−4	WB−5	WB−6	●	●	●	WB−7

Out-of-order execution addresses data hazards

- control hazards typically more expensive

(See also https://en.wikipedia.org/wiki/Instruction_pipelining)

ADM: Branch Misprediction & Predication

[interactive lecture (no slides; code examples in BrightSpace)]

ADM: Literature

- **Branch Misprediction & Predication**
  - “Conjunctive Selection Conditions in Main Memory”. Kenneth A. Ross. PODS 2002.
  - “Selection conditions in main memory”. Kenneth A. Ross. ACM Transactions On Database Systems 29: 132-161, 2004.

ADM: Agenda

- 07.09.2022: Lecture 1: **Introduction**
- 14.09.2022: Lecture 2: **SQL Recap**

  (plus Assignment 1 [in groups; 3 weeks]: TPC-H benchmark)
- 21.09.2022: Lecture 3: **Column-Oriented Database Systems (1/6) - Motivation & Basic Concepts**
- 28.09.2022: Lecture 4: **Column-Oriented Database Systems (2a/6) - Selected Execution Techniques (1/2)**

  (plus Assignment 2 [in groups; (1+)4 weeks]: Compression techniques)
- 05.10.2022: Lecture 5: **Column-Oriented Database Systems (2b/6) - Selected Execution Techniques (2/2)**
- 12.10.2022: Lecture 6: **Column-Oriented Database Systems (3/6) - Cache Conscious Joins**
- 19.10.2022: Lecture 7: **Column-Oriented Database Systems (4/6) - "Vectorized Execution"**
- ~~26.10.2022~~: ***No lecture!***
- 02.11.2022: Lecture 8: **Branch Misprediction & Predication**

  (plus Assignment 3 [individual; 2 weeks]: Predication)
- 09.11.2022: Lecture 9: **DuckDB: An embedded database for data science (1/2) (guest lecture & hands-on)**

  (plus Assignment 4 [individual; (1+)2 weeks]: Analysing NYC Cab dataset with DuckDB)
- 16.11.2022: Lecture 10: **DuckDB: An embedded database for data science (2/2) (guest lecture & *hands-on*)**
- 23.11.2022: Lecture 11: **Column-Oriented Database Systems (5/6) - Adaptive Indexing**
- 30.11.2022: Lecture 12: **Column-Oriented Database Systems (6/6) - Progressive Indexing**

Bring your own laptop!

ADM: Literature (6/6)

- **Column-Oriented Database Systems (5/6) - Adaptive Indexing**
  - “Cracking the Database Store”. Martin L. Kersten, Stefan Manegold. CIDR 2005.
  - “Database Cracking”. Stratos Idreos, Martin L. Kersten, Stefan Manegold. CIDR 2007.
  - “Self-selecting, self-tuning, incrementally optimized indexes”. Goetz Graefe, Harumi A. Kuno. EDBT 2010.
  - “Merging What's Cracked, Cracking What's Merged: Adaptive Indexing in Main-Memory Column-Stores”. Stratos Idreos, Stefan Manegold, Harumi A. Kuno, Goetz Graefe. Proc. VLDB Endow. 4(9): 585-597, 2011.
  - “Stochastic Database Cracking: Towards Robust Adaptive Indexing in Main-Memory Column-Stores”. Felix Halim, Stratos Idreos, Panagiotis Karras, Roland H. C. Yap. Proc. VLDB Endow. 5(6): 502-513, 2012.