Session V-A: Domain Specific Accelerators (Tuesday 1:30PM)

SMASH

Co-designing Software Compression and Hardware-Accelerated Indexing for Efficient Sparse Matrix Operations

Konstantinos Kanellopoulos, Nandita Vijaykumar, Christina Giannoula,
Roknoddin Azizi, Skanda Koppula, Nika Mansouri Ghiasi,
Taha Shahroodi, Juan Gomez Luna, Onur Mutlu

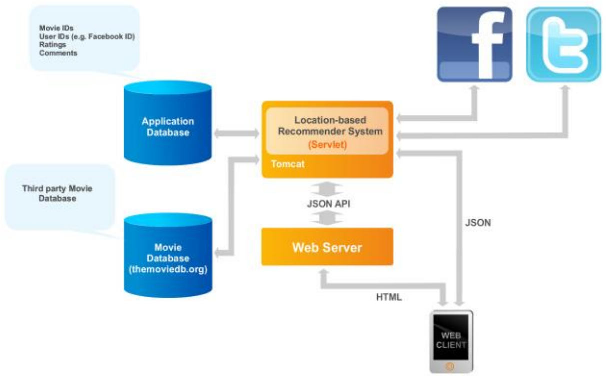
SAFARI

ETH zürich

Carnegie Mellon University

Sparse Matrix Operations are Widespread Today

Recommender Systems


Movie IDs
User IDs (e.g. Facebook ID)
Ratings
Comments
Application Database
Location-based Recommender System
(Servlet)
Tomcat
Third party Movie Database
JSON API
JSON
Movie Database (themoviedb.org)
Web Server
HTML
WEB CLIENT

- Collaborative Filtering

Graph Analytics

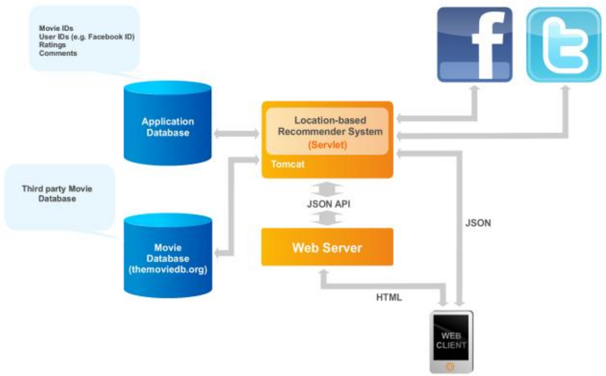
- PageRank
- Breadth-First Search
- Betweenness Centrality

Neural Networks

- Graph Neural Networks
- Sparse Deep Neural Networks

Sparse Matrix Operations are Widespread Today

Recommender Systems


Movie IDs
User IDs (e.g. Facebook ID)
Ratings
Comments
Application Database
Third party Movie Database
Movie Database (themoviedb.org)
Location-based Recommender System
(Servlet)
Tomcat
JSON API
Web Server
JSON
HTML
WEB CLIENT

- Collaborative Filtering

Graph Analytics

- PageRank
- Breadth-First Search
- Betweenness Centrality

Neural Networks

- Graph Neural Networks
- Sparse Deep Neural Networks

Sparse matrix compression is essential to enable efficient storage and computation

Limitations of Existing Compression Formats

Limitations of Existing Compression Formats

❶

General formats optimize for storage → **Expensive discovery of the positions of non-zero elements**

Limitations of Existing Compression Formats

❶

General formats optimize for storage → **Expensive discovery of the positions of non-zero elements**

❷

Specialized formats assume specific matrix structures and patterns (e.g., diagonals) → **Narrow applicability**

SMASH

SMASH

Hardware/Software cooperative mechanism:

- Enables **highly-efficient** sparse matrix compression and computation
- **General** across a diverse set of sparse matrices and sparse matrix operations

SMASH

Hardware/Software cooperative mechanism:

- Enables **highly-efficient** sparse matrix compression and computation
- **General** across a diverse set of sparse matrices and sparse matrix operations

Software

Efficient compression using a Hierarchy of Bitmaps

SMASH

Hardware/Software cooperative mechanism:

- Enables **highly-efficient** sparse matrix compression and computation
- **General** across a diverse set of sparse matrices and sparse matrix operations

Software

Efficient compression using a Hierarchy of Bitmaps

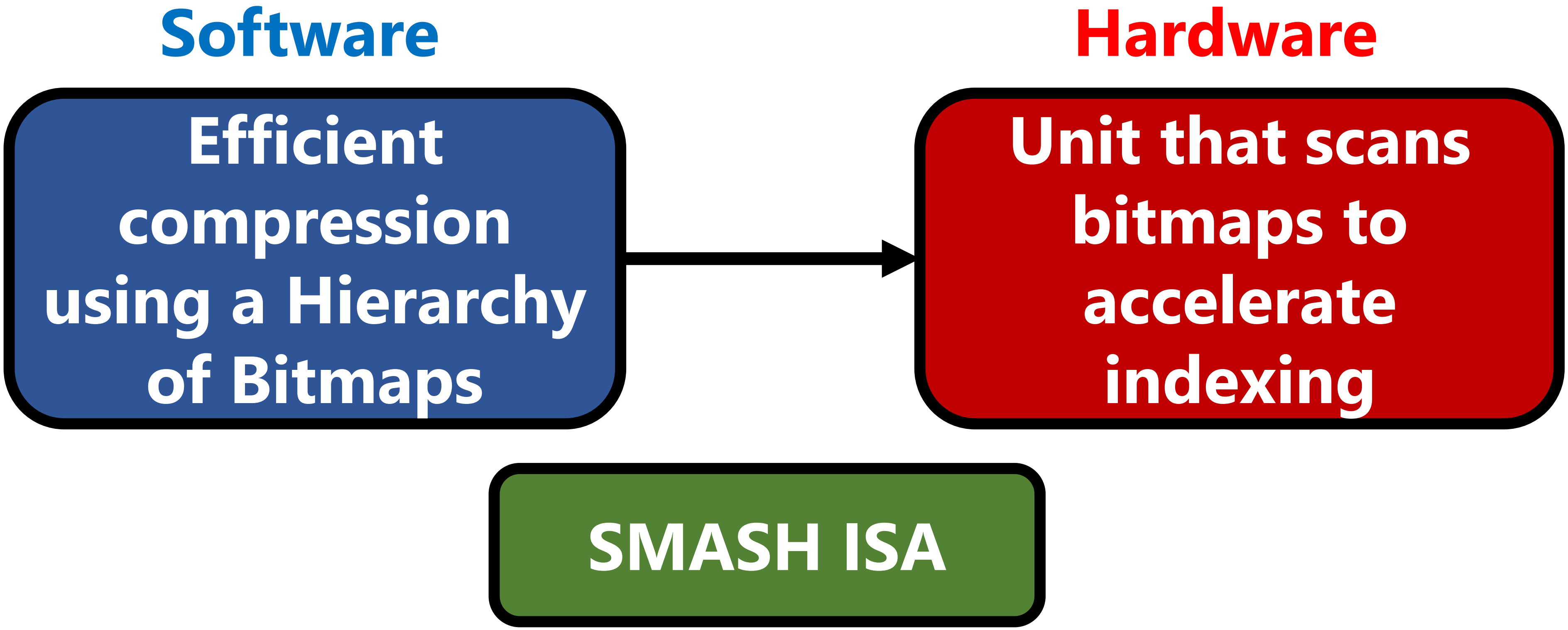
Hardware

Unit that scans bitmaps to accelerate indexing

SMASH

Hardware/Software cooperative mechanism:

- Enables **highly-efficient** sparse matrix compression and computation
- **General** across a diverse set of sparse matrices and sparse matrix operations



Software

Efficient compression using a Hierarchy of Bitmaps

Hardware

Unit that scans bitmaps to accelerate indexing

SMASH ISA

Key Results

SMASH

- 38% and 44% speedup for SpMV and SpMM

Hardware Overhead

- 0.076% area overhead over an Intel Xeon CPU

Session V-A: Domain Specific Accelerators (Tuesday 1:30PM)

SMASH

Co-designing Software Compression and Hardware-Accelerated Indexing for Efficient Sparse Matrix Operations

Konstantinos Kanellopoulos, Nandita Vijaykumar, Christina Giannoula, Roknoddin Azizi, Skanda Koppula, Nika Mansouri Ghiasi, Taha Shahroodi, Juan Gomez Luna, Onur Mutlu

SAFARI ETH zürich Carnegie Mellon University