DBs and hardware: the beginning and sequel of a beautiful friendship

Anastasia Ailamaki

EPFL and RAW Labs SA
FRIEND OR BULLY?

pipelining
ILP
SMT
multi-core
heterogeneous H/W
recurring headaches

non-uniform communication

underutilization of resources

one size fits none

aspirin or vitamin?
recurring headaches

non-uniform communication

underutilization of resources

one size fits none

aspirin or vitamin?
NU: MA, and more

unpredictable performance
scalability on multicores

OLTP
8 socket x 10 core
800K rows
Probing one row

OLAP
2 socket x 8 core
Parallel aggregation
128MB column

vitamin: adaptive partitioning
OLTP: communication islands

Throughput (MTPS)

Number of sockets

near-linear scalability

[VLDB12, ICDE14]
OLAP: partitioning across sockets

 através de sockets

 1 hot table

 3 hot tables

Throughput (QphH)

Throughput (nav. steps/h)

# of partitions per table

 fewer partitions ➔ higher throughput

[VLDB15]
recurring headaches

non-uniform communication

underutilization of resources

one size fits none

aspirin or vitamin?
processor is idle most of time

use cores
improve performance
save energy

TPC-C
100GB data
1 worker thread
Intel Ivy Bridge

STM
DBMS
VoltDB
Hyper
DBMS

100x
instruction stalls (still) rule

100GB data on Shore-MT
Intel Xeon E5-2660

[EDBT13,DAMON13]

vitamin: follow L1-I contents
SLICC: chasing locality in L1-I

[MICRO12, ISCA13, VLDB15]

improved performance by 70%
recurring headaches

non-uniform communication

underutilization of resources

one size fits none

aspirin or vitamin?
one DBMS fits all

but: data and apps vary, so...

→ one DBMS fits one app/format/...
one DBMS fits NONE
access only useful raw data (not clear which data)
no copying or altering data
no vendor lock-in
custom (per-query) engine

vitamin: just-in-time databases

RAWdb: an evolving query engine
how to build a just-in-time database

```
SELECT event.jet...
FROM csv, root
WHERE csv.RunNumber == root.RunNumber AND
root. EF_2mu13 == TRUE AND ...
```

Code Generate the Access Paths

Code Generate the Query

Build Position and Data Caches

RAWdb is invisible
adapts to data and queries
aspirin → lag behind

catch me

if you can
adaptive data partitioning
chasing instruction locality
just-in-time databases
BACKUP
Databases and Hardware

• abundant compute power
  – increasing heterogeneous parallelism
  – deep storage hierarchies

• to use it
  – adapt to hardware and application
  – Just-in-Time-Time Databases
adding compute elements

2005

2020

Core

pipelining

ILP

SMT

multisocket multicores

heterogeneous many-cores

abundant heterogeneous parallelism
storage depths

Latency

Bandwidth

depth storage hierarchy
once upon a time ...

processor stalled 50% of time
non-uniform communication latencies

- Threads: <10 cycles
- Inter-socket links: 50 cycles
- Inter-socket links: 500 cycles

Socket 0

Socket 1

non-uniform
How to build a just-in-time data base

SELECT event.jet...
FROM csv, root
WHERE csv.RunNumber == root.RunNumber AND
  root.EF_2mu13 == TRUE AND ...

... containing physics events

... containing “good” run numbers
Optimize just-in-time for any application.

- **Application**
  - OLTP
  - Analytics
  - Streams
  - ML

- **Data Storage**
  - Rowstore
  - Columnstore
  - Hybrid
  - Hierarchical

- **HW platform**
  - Multisockets
  - GPUs
  - Accelerators
  - Rackscale
obs: overlap across transactions

TPC-C (100GB data) on Shore-MT
Overlapping cache blocks

higher overlap among same-type transactions
advanced instruction chasing

**ADDICT**

**Phase 1: Profile**
- Threads: T1, T2
- L1I
- start
- probe1
- probe2
- update

**Phase 2: Migrate**
- CORES
- start
- probe1
- probe2
- update

compared to SLICC: minimal changes to hw & 2x higher performance improvement
Perfectly partitionable workload

Island awareness brings scalability

Throughput (MTPS)

Number of sockets

Shared-nothing
ATraPos
PLP

8 socket x 10 core
800K row dataset
Probing one row
• Try harder to maximize use of hardware
• make databases invisible
ATraPos: adaptive transaction processing

adaptive partitioning and placement
adapting to islands

Throughput (MTTPS)

Number of sockets

near-perfect scalability

Shared-nothing
ATraPos
PLP

8 socket x 10 core
800K row dataset
Probing one row
one DBMS fits one app/data/...

Main-memory DBMS

Column stores

NoSQL systems

Stream DBMS

OLTP

OLAP

Large-scale embarrassingly parallel