The Glass Half Full

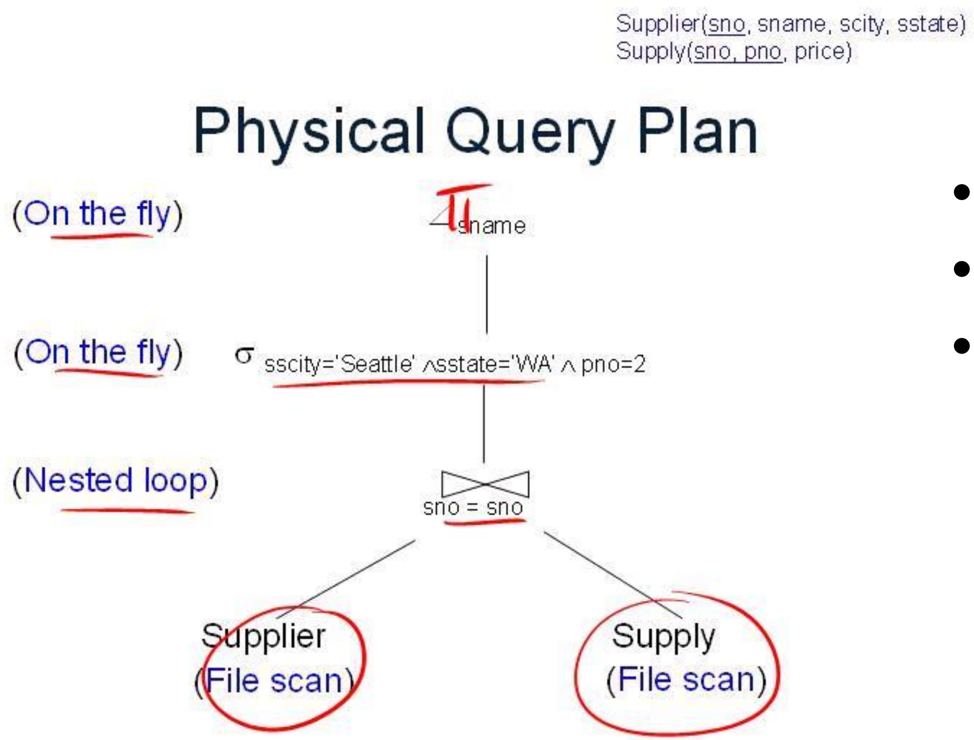
Using Programmable Hardware Accelerators in Analytical Databases

Zsolt István

IMDEA Software Institute

Context: Database Query Execution

"List all suppliers from Seattle, WA who have part no. 2 in stock"



- Streaming operators
- Can "push down" filters
- Can split plan across devices

Context: Analytical Databases

- OLAP – Online Analytical Processing
  - Large datasets – up to TBs
  - Ad-hoc querying to extract insight, recurring reporting – Possibly complex operations
  - Read-mostly workloads, updates in batches

- OLTP – Online Transaction Processing
  - Smaller datasets
  - Queries known, relate to business actions
  - Makes heavy use of indexes
  - Reads and updates intermixed

Databases were a 25 Billion $ market in 2018, much of it happening in the cloud…

Could we specialize machines to them?

https://www.statista.com/statistics/810188/worldwide-commercial-database-market-size/

Database Computer – '70s

"The first goal is to design it with the capability of handling a very large on-line database of 10^10 bytes or beyond since special-purpose machines are not likely to be cost-effective for small databases."

- Fully custom machine for databases
  - Processors – special ISA microprocessors
  - Memory – magnetic bubbles and CCDs
- Semiconductor technology and general purpose CPUs took over

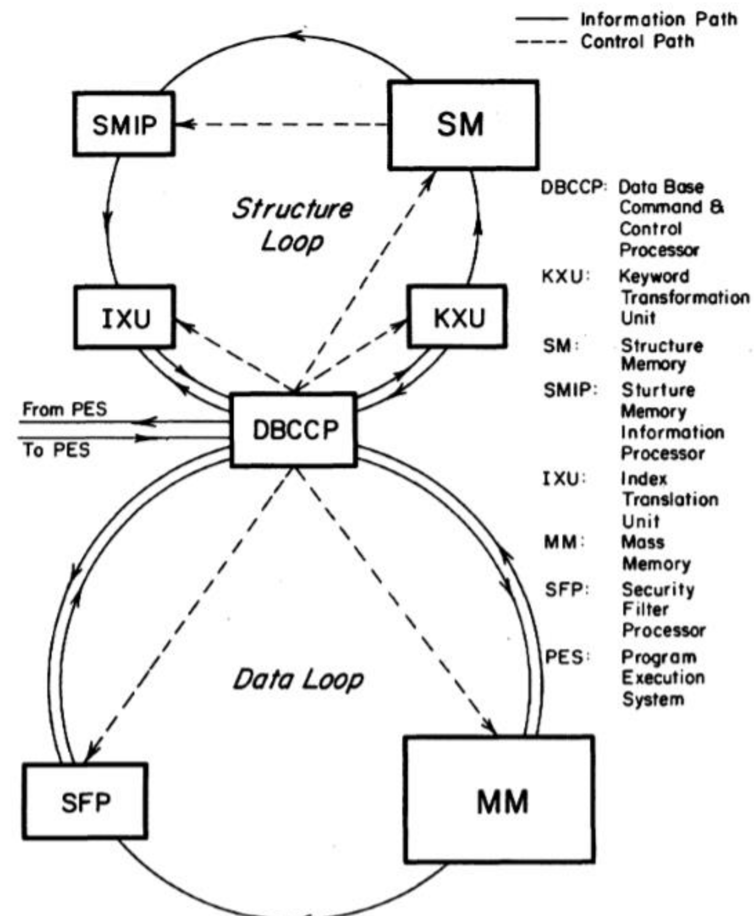


Fig. 1. Architecture of the DBC.

Jayanta Banerjee, David K. Hsiao, Krishnamurthi Kannan: *DBC - A Database Computer for Very Large Databases*. IEEE Trans. Computers 28(6): 414-429 (1979)

Gamma Machine – ’80s

- Based on VAX multi-processor system
- By the time the software and hardware were developed, CPUs have become much faster
    - Couldn’t keep up with Moore’s law

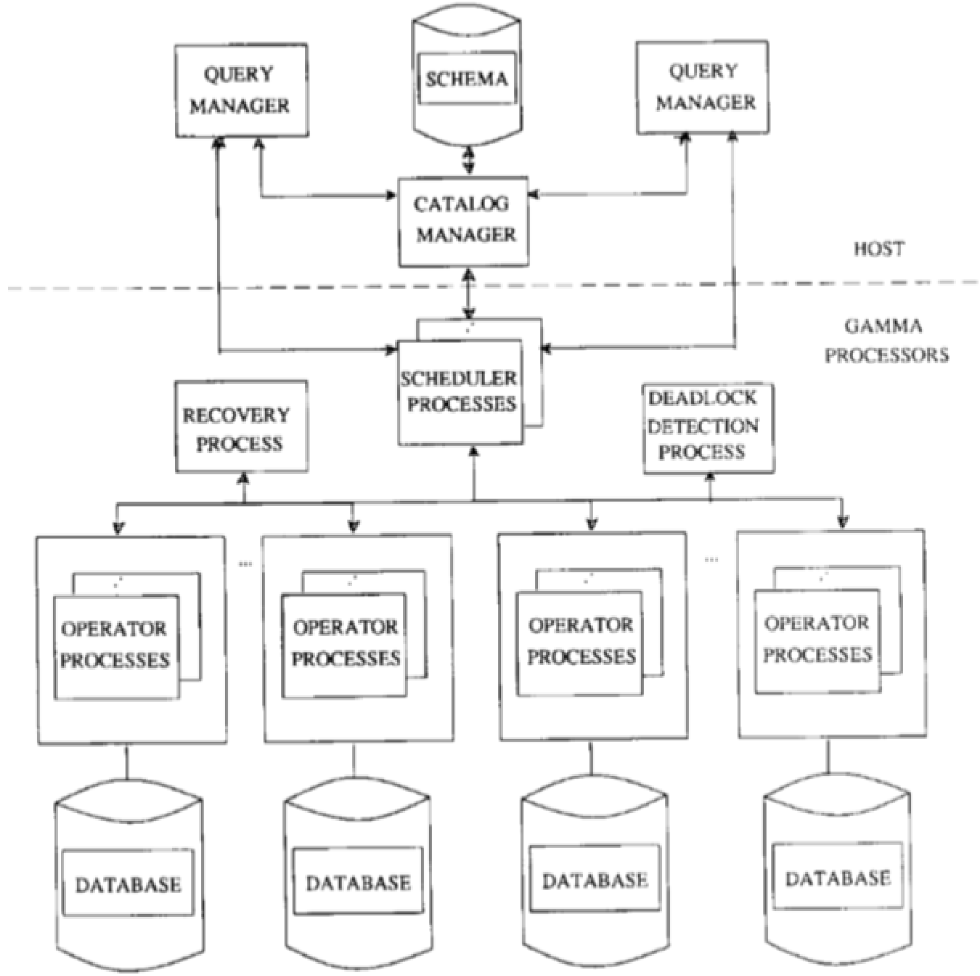
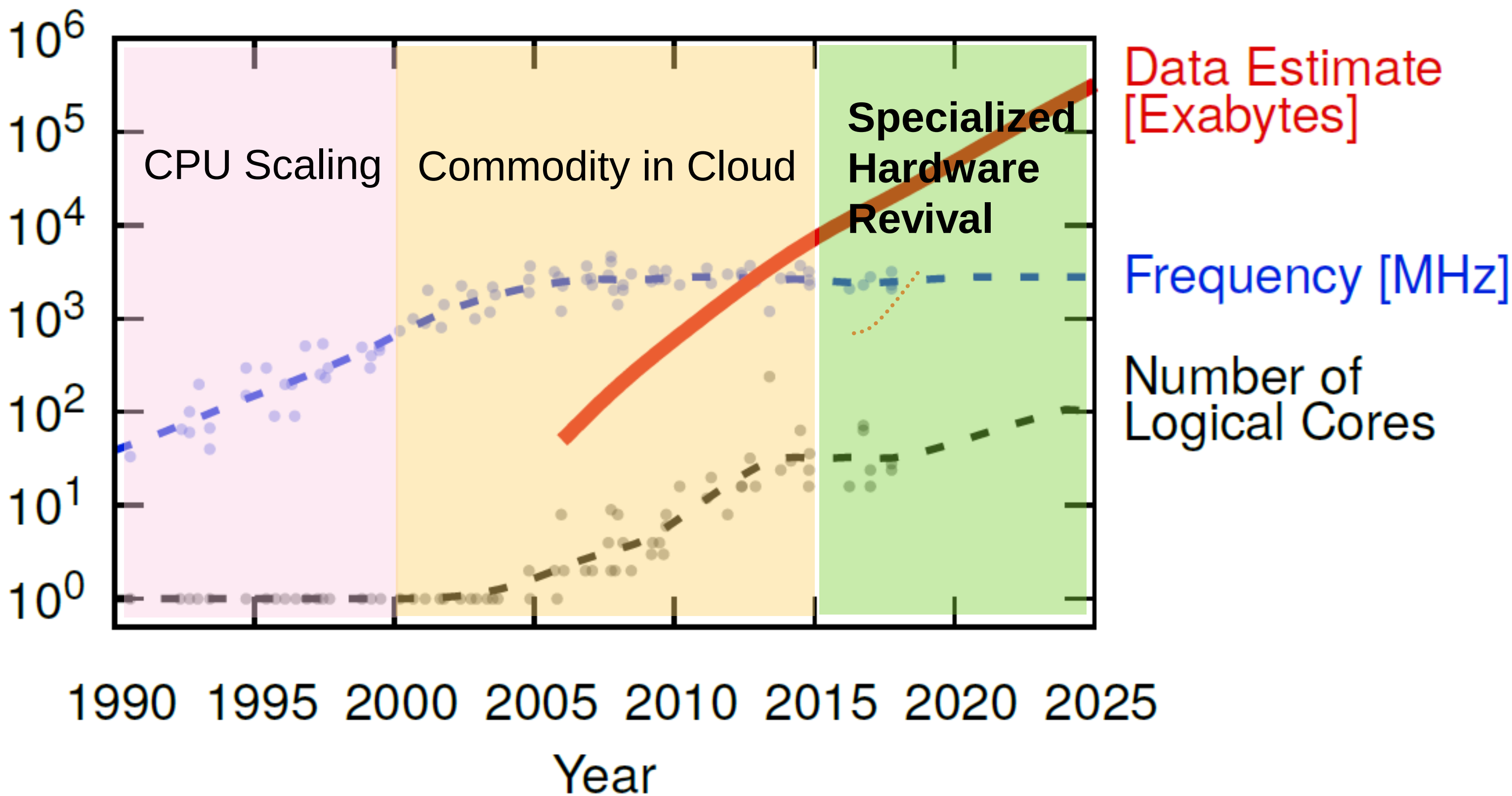


Fig. 2. Gamma process structure.

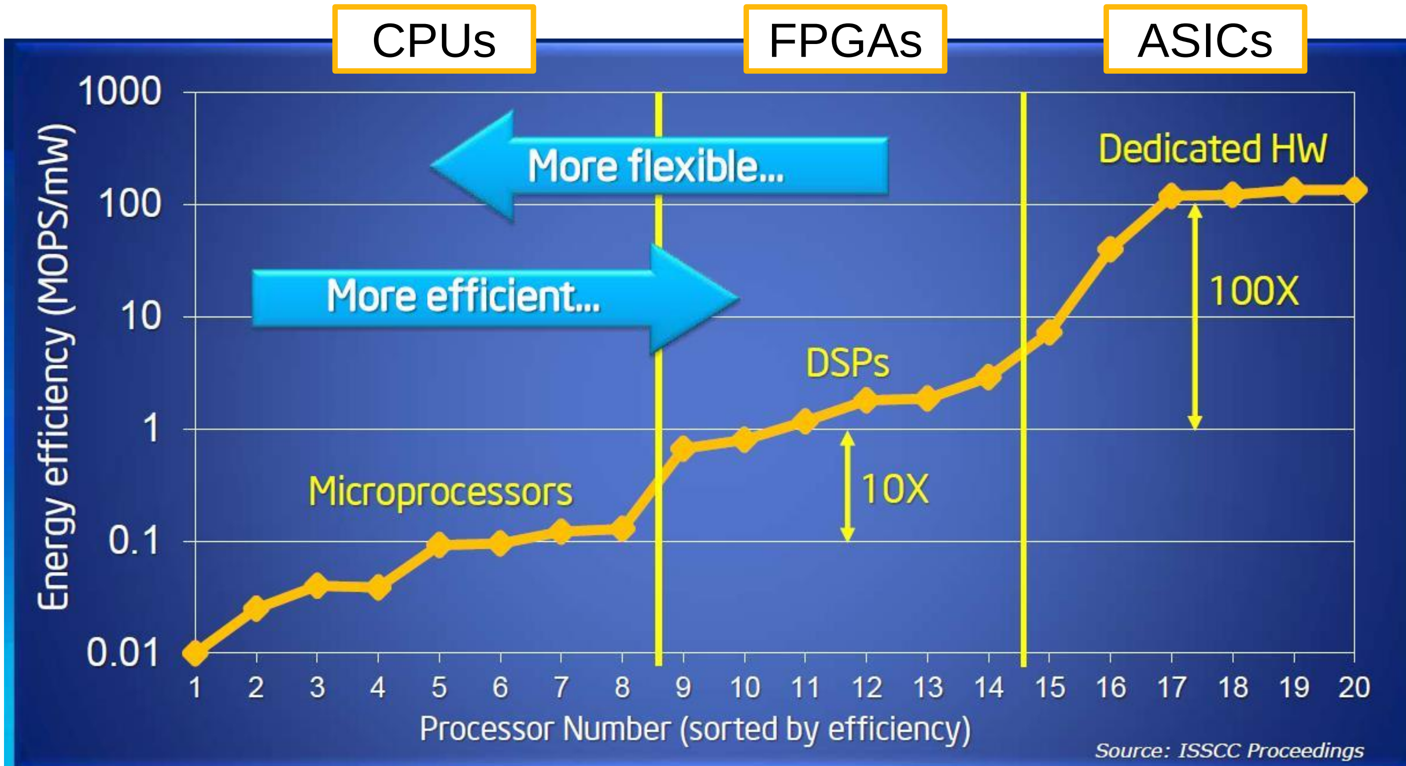
David J. DeWitt, Robert H. Gerber, Goetz Graefe, Michael L. Heytens, Krishna B. Kumar, M. Muralikrishna: *GAMMA - A High Performance Dataflow Database Machine.* VLDB 1986: 228-237

Data/Compute Gap

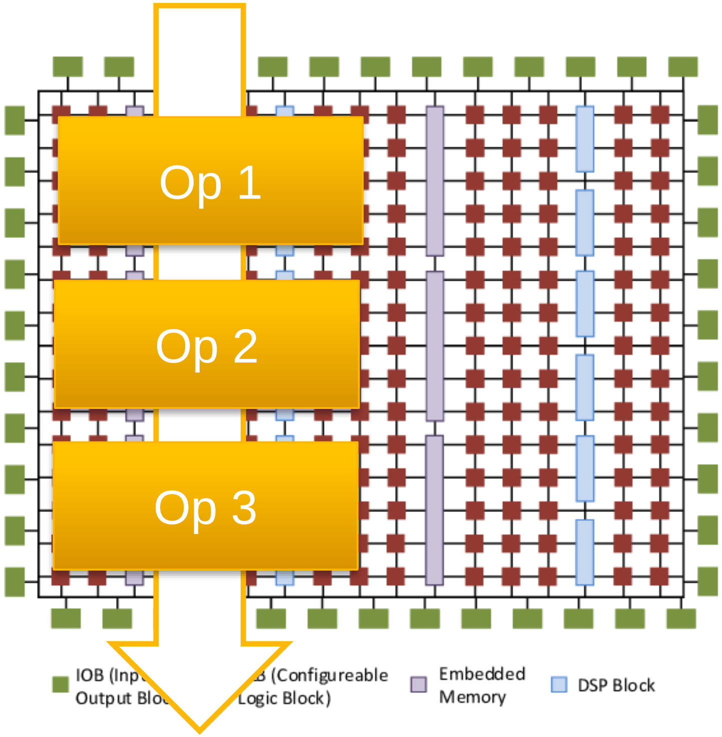


Based on a plot layout by K. Rupp. Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten; 2010-2015 by K. Rupp. Data growth estimate by C. Maxfield.

Renewed interest in Specialized Hardware



Re-programmable Specialized Hardware

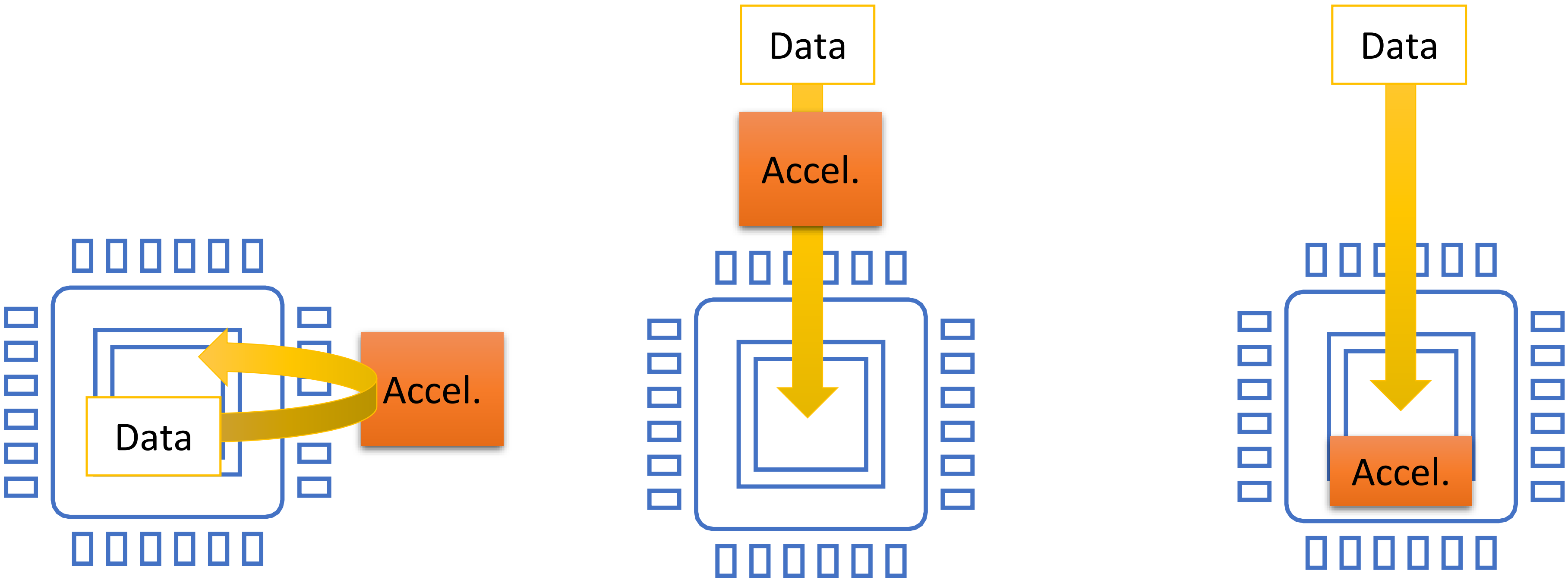


Field **P**rogrammable **G**ate **A**rray (FPGA)

- Free choice of architecture
- Fine-grained pipelining, communication, distributed memory
- Tradeoff: all “code” occupies chip space

- Evolving platform: larger chips, more heterogeneity

Integration Options



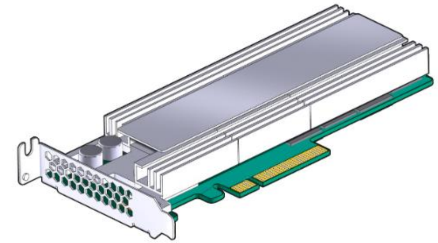
1) On the side

2) In data-path

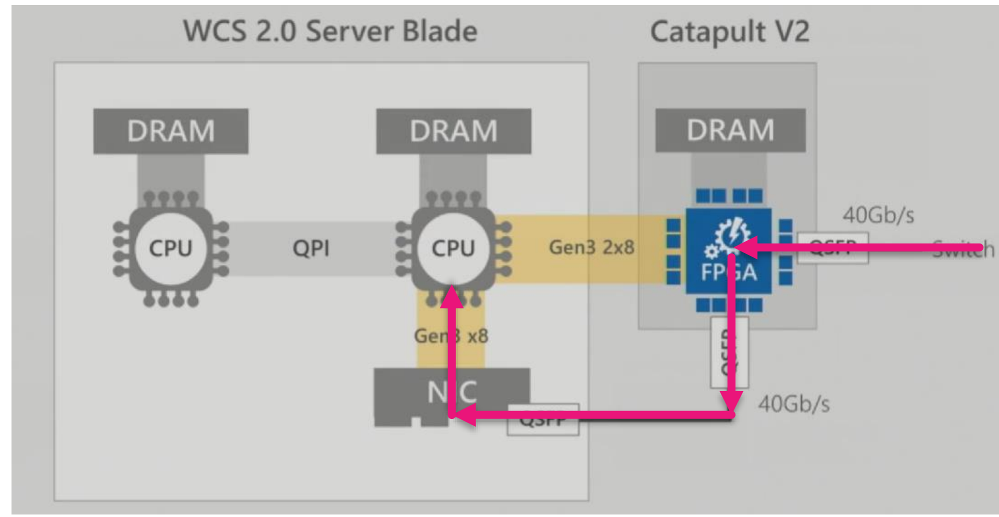
3) Co-processor

In the Cloud Today

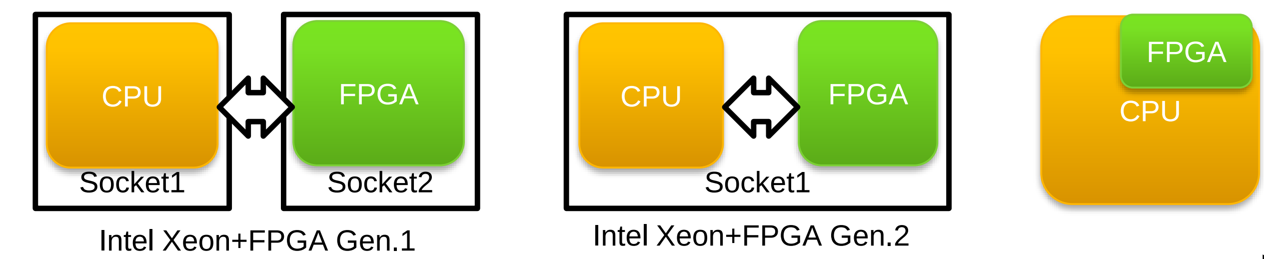
- Accelerator
  - Amazon F1



- In data path
  - Microsoft Catapult



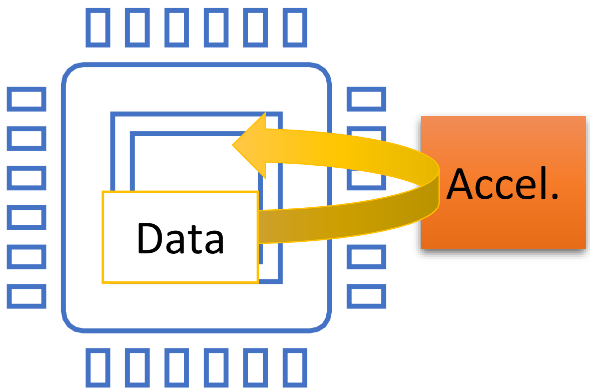
- Co-processor
  - Intel Xeon+FPGA

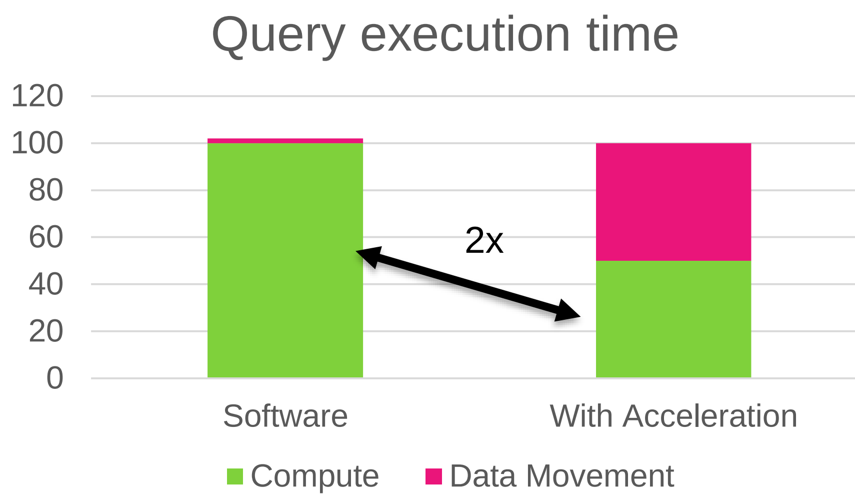


The Glass Half Empty…

The Glass Half Empty…

- 1) On the side acceleration introduces overhead

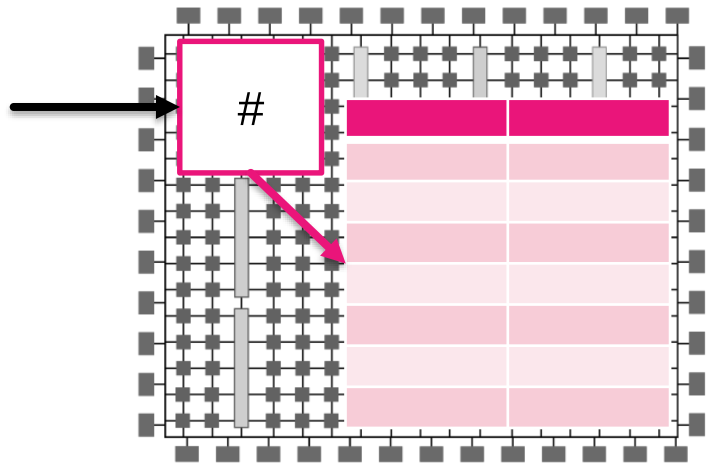




- Many related work offers no real speedup if we factor in data movement, transformation, software overhead…

The Glass Half Empty…

- 2) "All or nothing" behavior makes query planning difficult
  - Example: fixed capacity hash table on FPGA
  - Constant time access for reads and writes
  - What happens if data doesn't fit?
    - Can't always know the number of keys aprioi



The Glass Half Empty…

- 3) Analytical databases becoming more optimized / not much compute in core SQL
- X100 [CIDR05] showed that <10% of compute time spent on SQL operators +,-,*,SUM,AVG in analytical queries
  - Columnar stores often memory bound (10s of GB/s)

RowId	EmpId	Lastname	Firstname	Salary
001	10	Smith	Joe	40000
002	12	Jones	Mary	50000
003	11	Johnson	Cathy	44000
004	22	Jones	Bob	55000

The Glass Half Empty…

- On the side acceleration introduces overhead
- “All or nothing” behavior makes query planning difficult
- Analytical databases becoming more optimized / not much compute in core SQL

The Glass Half Full…

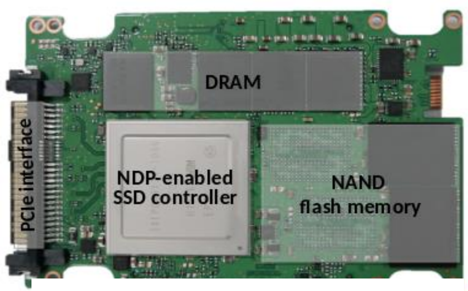
- ~~On the side acceleration introduces overhead~~
- ✓ Reduce data movement bottlenecks

Processing in data path: Smart Flash

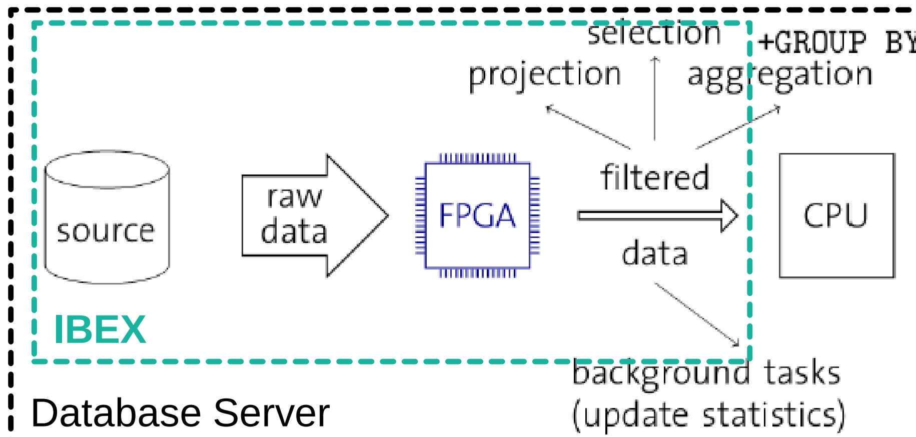
- IBEX: Database storage engine with processing offload
  - Filter and pre-aggregate for analytic workloads

→ Larger bandwidth, more IOPS (Samsung YourSQL, MIT BlueDBM)

- Opportunity to extend SSDs/Flash with complex offload



Samsung "smart" SSD



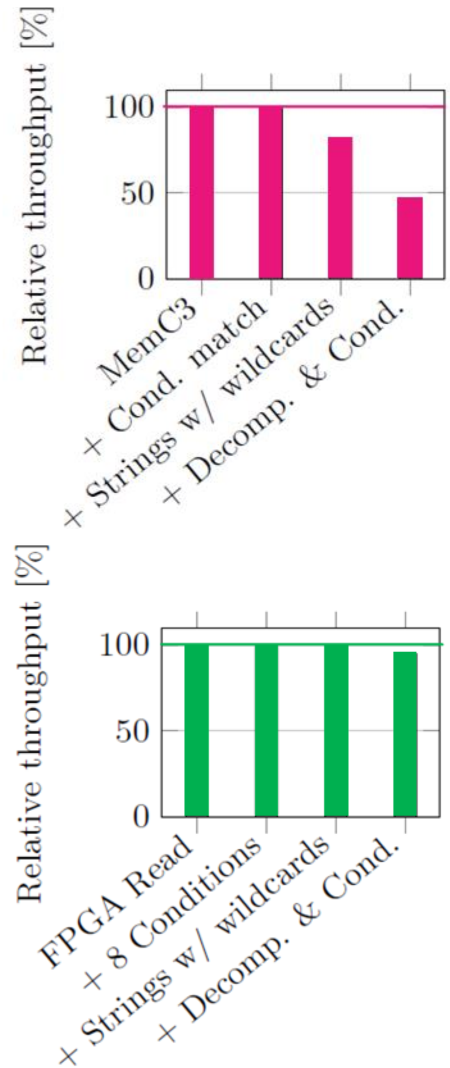
IBEX – An Intelligent Storage Engine with Support for Advanced SQL Off-loading. L. Woods, Z. Istvan and G. Alonso, VLDB'14

Smart Storage in Databases: Filter push-down

```
SELECT … FROM customer
WHERE  age<35 AND purchases>2
AND address LIKE “%PO. Box 123%”
```

- **Challenge: guarantee that filtering never slows down retrieval**

- In HW algorithms can be re-designed to become bandwidth-bound instead of compute-bound

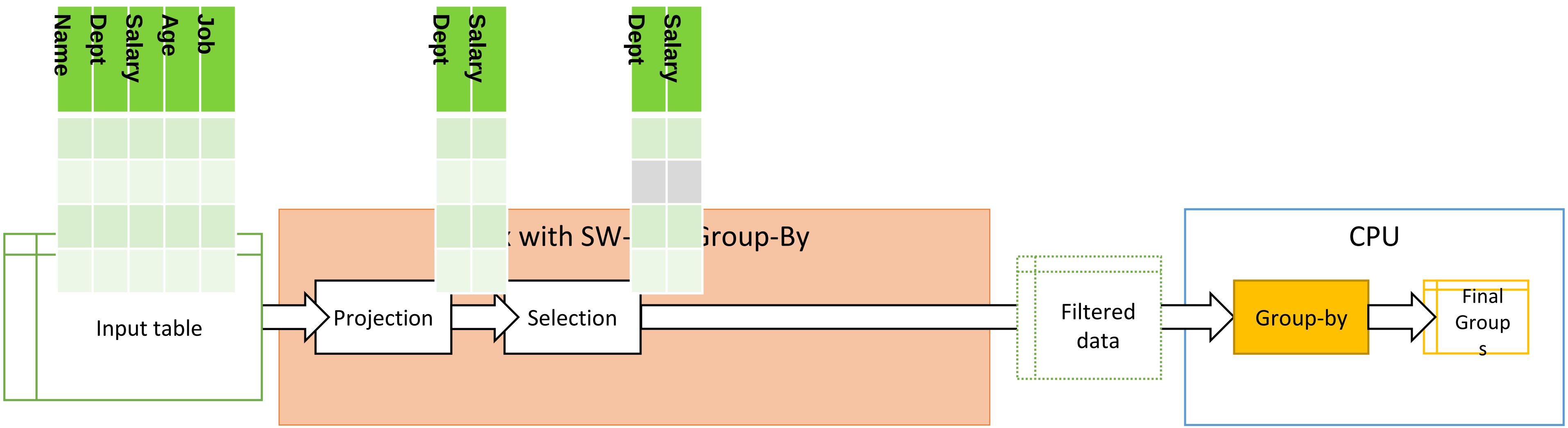


The Glass Half Full…

- ✓ Reduce data movement bottlenecks
- ~~“All or nothing” behavior makes query planning difficult~~
- ✓ Hybrid processing

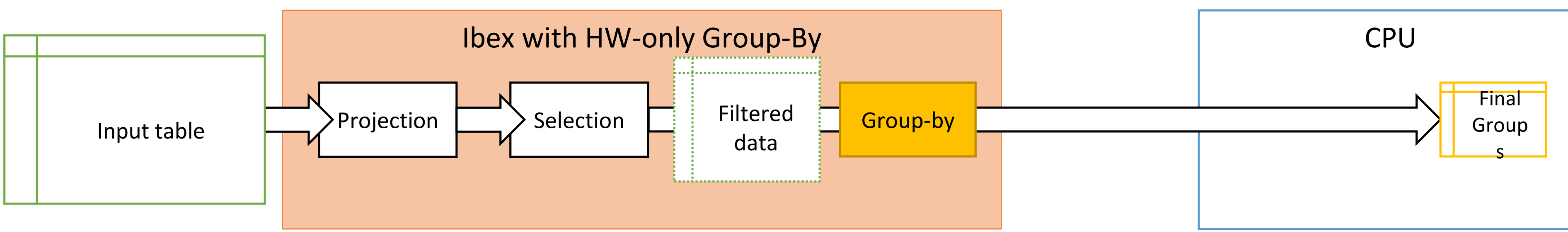
IBEX's Hybrid Group-by

- Group-by: Compute aggregate function over categories
  - `select avg(salary) from employees group by department`



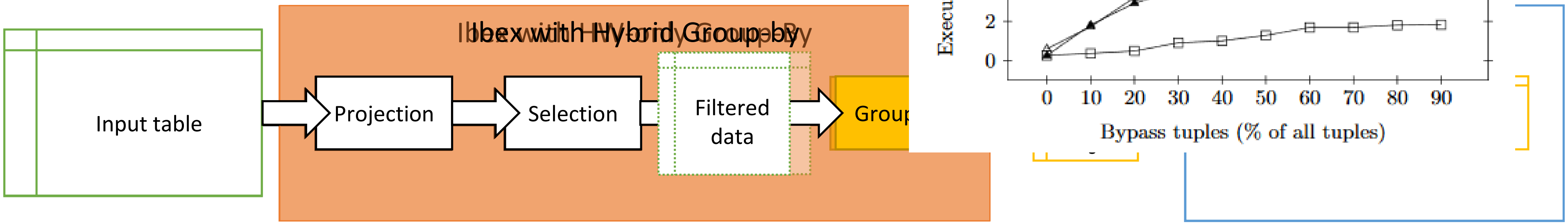
IBEX's Hybrid Group-by

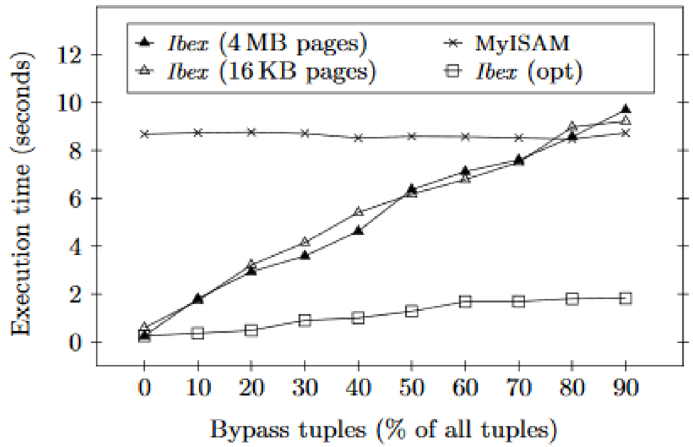
- Group-by: Compute aggregate function over categories
  - `select avg(salary) from employees group by department`



IBEX's Hybrid Group-by

- Group-by: Compute aggregate function over categories
  - `select avg(salary) from employees group by department`
- If number of groups does not fit on FPGA?
  - Send partial aggregates – finalize in SW
  - Worst case: same as no acceleration
  - Best- case: All in HW!





Challenge: How to split across accelerator and software?

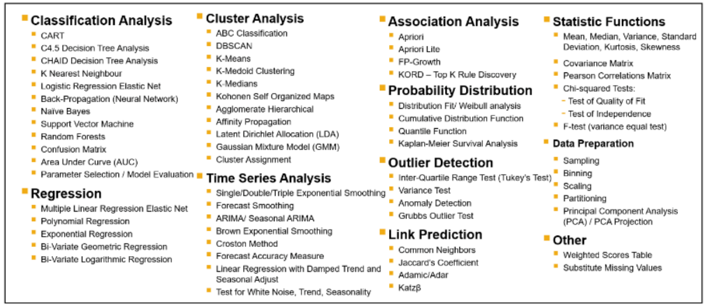
The Glass Half Full

- ✓ Reduce data movement bottlenecks
- ✓ Hybrid Processing

- ~~Analytical databases becoming more optimized / not much compute in core SQL~~
- ✓ Emerging compute-intensive workloads

The Rise of Machine Learning

- Databases adopting new ways of analyzing the data
  - SAP Hana, Oracle, SQL Server, etc.



- Specialized hardware can help both with model building and inference
- Benefits for “classical” algorithms as well

The Glass Half Full

- ✓ Reduce data movement bottlenecks
- ✓ Hybrid Processing
- ✓ Emerging compute-intensive workloads

- Are we done?

Compilation/synthesis of hardware accelerators

- Can we derive accelerators directly from user queries or we design a set of "specialized cores"?

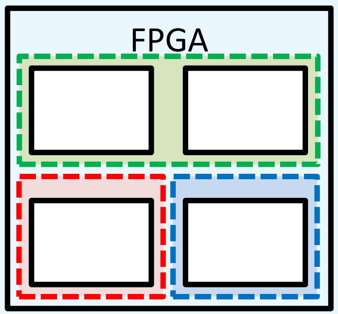
- Intermediary DSL to make compilation easier/feasible?
  - Many operators share building blocks, e.g., data hashing, sorting, shuffle, etc.
  - Can we assemble circuits from these coarser grained components?

Managing programmable hardware accelerators

- Is the accelerator managed and configured by the OS/Hypervisor, or does the Database take control?

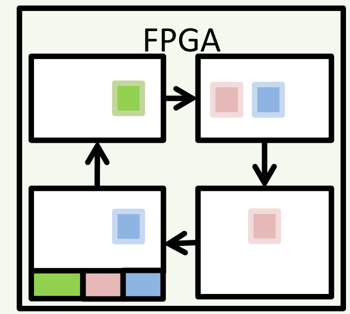
- **How to share programmable hardware across tenants?**
  - Data isolation
  - Performance isolation

Providing multi-tenancy with FPGAs



Virtualization

- General purpose (PR)
- Few tenants
- Trades off functionality
- Course grained resource alloc.
- Tenants “bring” applications



Multi-tenant applications

- Domain-specific
- Many tenants
- Trades off performance (?)
- Fine grained resource alloc.
- Provider “brings” application

Providing Multi-tenant Services with FPGAs: Case Study on a Key-Value Store. Zs. István, G. Alonso. A. Singla. *28th International Conference on Field Programmable Logic and Applications (FPL'18), Dublin, Ireland, August 2018.*

The Glass Half Full…

- ✓ Reduce data movement bottlenecks
- ✓ Hybrid Processing
- ✓ Emerging compute-intensive workloads

Future Challenges…

- ? Managing programmable hardware accelerators
- ? Compilation/synthesis of hardware accelerators

For more details, see*: The Glass Half Full: Using Programmable Hardware Accelerators in Analytics.* Z. István. IEEE Data Engineering Bulletin, March 2019.