The Data Cyclotron

Romulo Goncalves
Martin Kersten

Distributed query processing

- SOLD OUT ON IDEAS?

MAP-REDUCE IS THE
HORSE POWER YOU NEED

CLOUDS OBSCURE YOUR VISION
AND KILL YOUR HOLIDAY PLEASURE

"Thinking Outside the Box."

- The holy grail of distributed query processing II
  - Organic growing scalable architecture
  - Crowd coordination rather then masters' control
  - Nothing remains the same, turbulent Data
  - Continuous Self-organization

THE UNIVERSE OF DB ARCHITECTURES IS SPARSELY EXPLORED !!!!

A classical design issue

Move the computation to the data, because

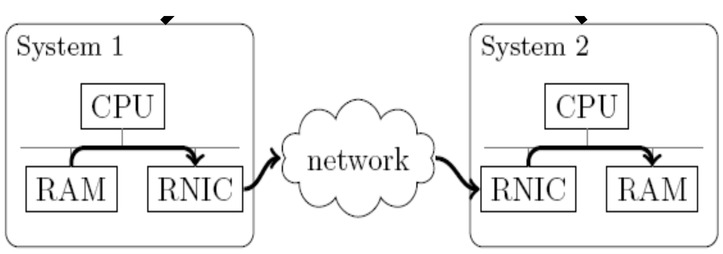
data transport is expensive

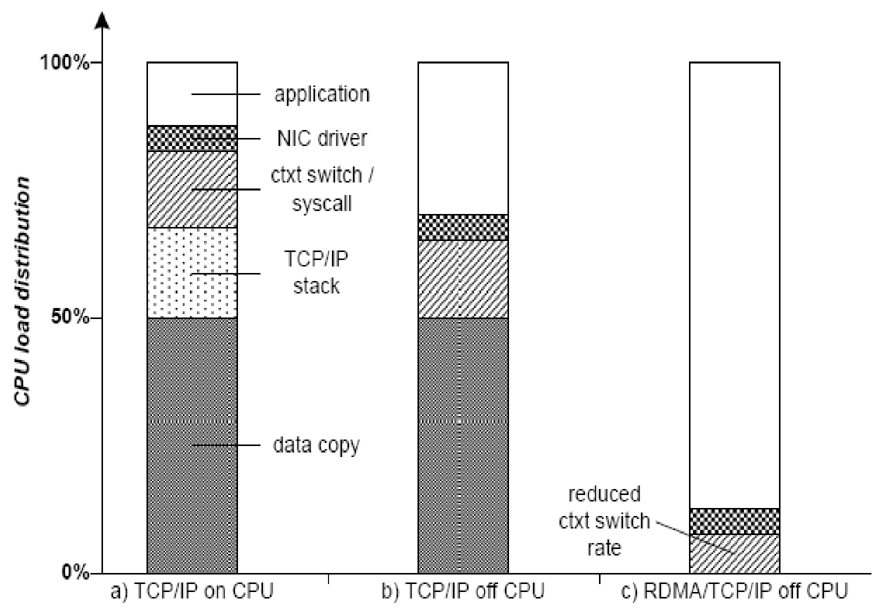
Hypothesis:
Make transport an asset rather then a problem

Remote Direct Memory Access

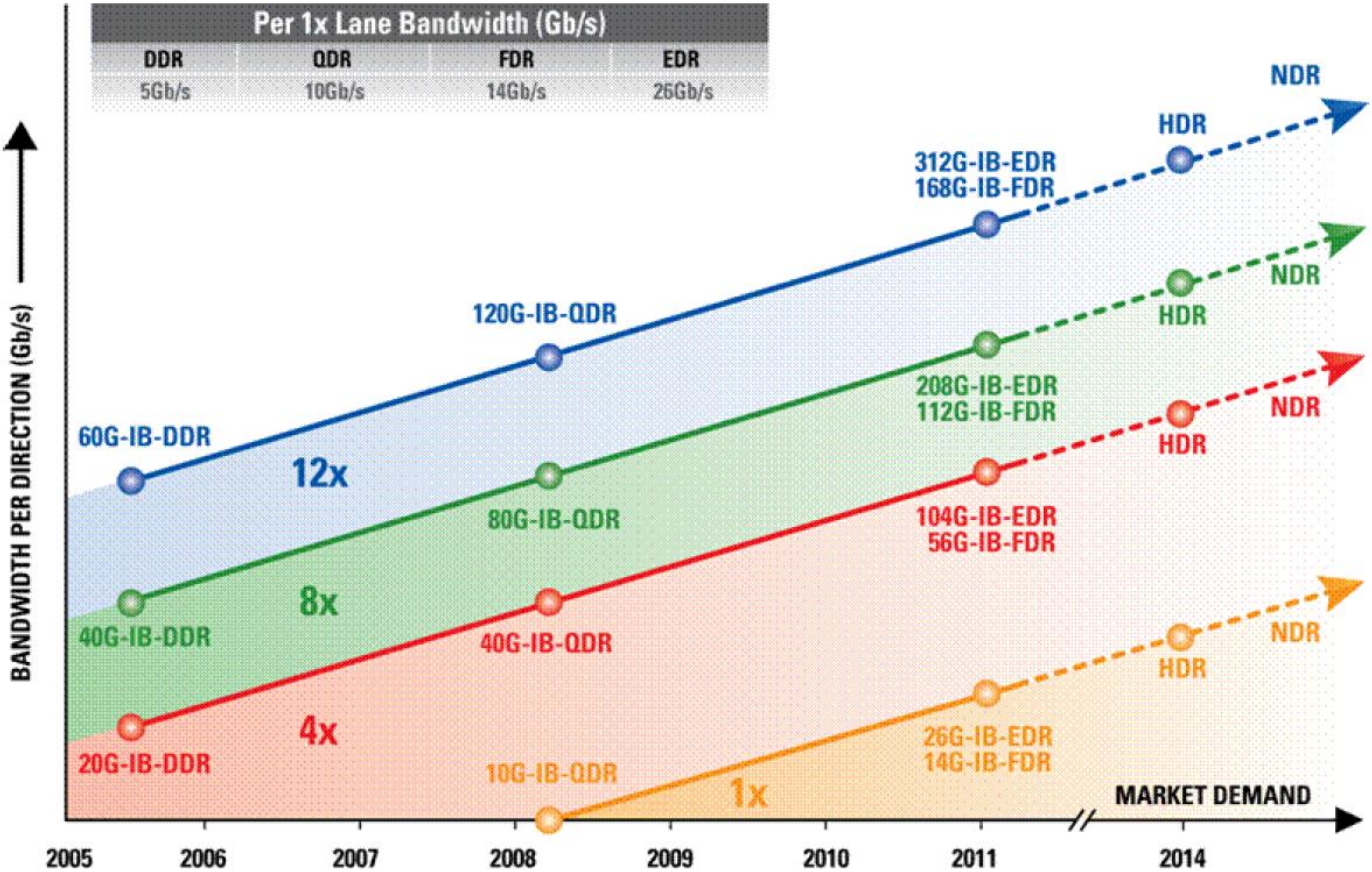
Remote Direct Memory Access (RDMA)

- Remote Memory at Your Finger Tips..
  - Significant reduced CPU load
  - Reduced Memory Bus Traffic





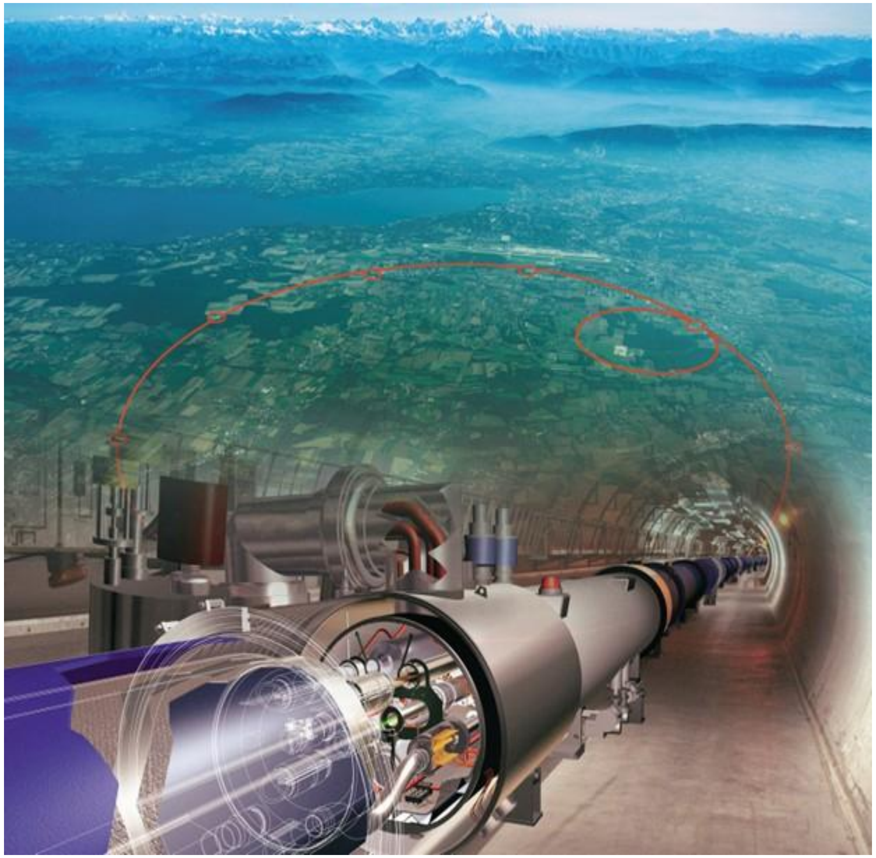
CWI



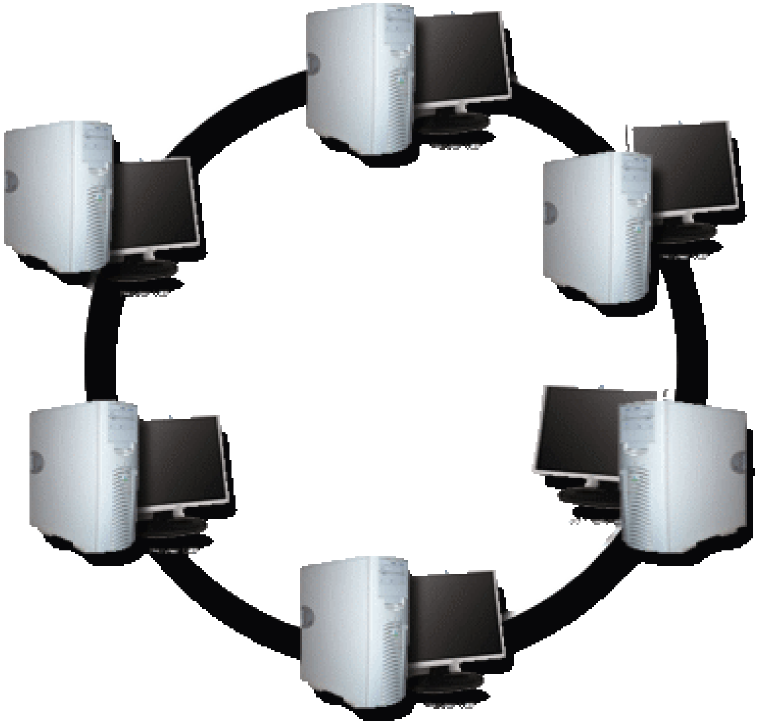
Per 1x Lane Bandwidth (Gb/s)
DDR
5Gb/s
QDR
10Gb/s
FDR
14Gb/s
EDR
26Gb/s
BANDWIDTH PER DIRECTION (Gb/s)
60G-IB-DDR
40G-IB-DDR
20G-IB-DDR
12x
8x
4x
1x
120G-IB-QDR
80G-IB-QDR
40G-IB-QDR
10G-IB-QDR
312G-IB-EDR
168G-IB-FDR
208G-IB-EDR
112G-IB-FDR
104G-IB-EDR
56G-IB-FDR
26G-IB-EDR
14G-IB-FDR
HDR
NDR
HDR
NDR
HDR
NDR
HDR
NDR
MARKET DEMAND
2005
2006
2007
2008
2009
2010
2011
2014

The topology.

- Swiss one (LHC)

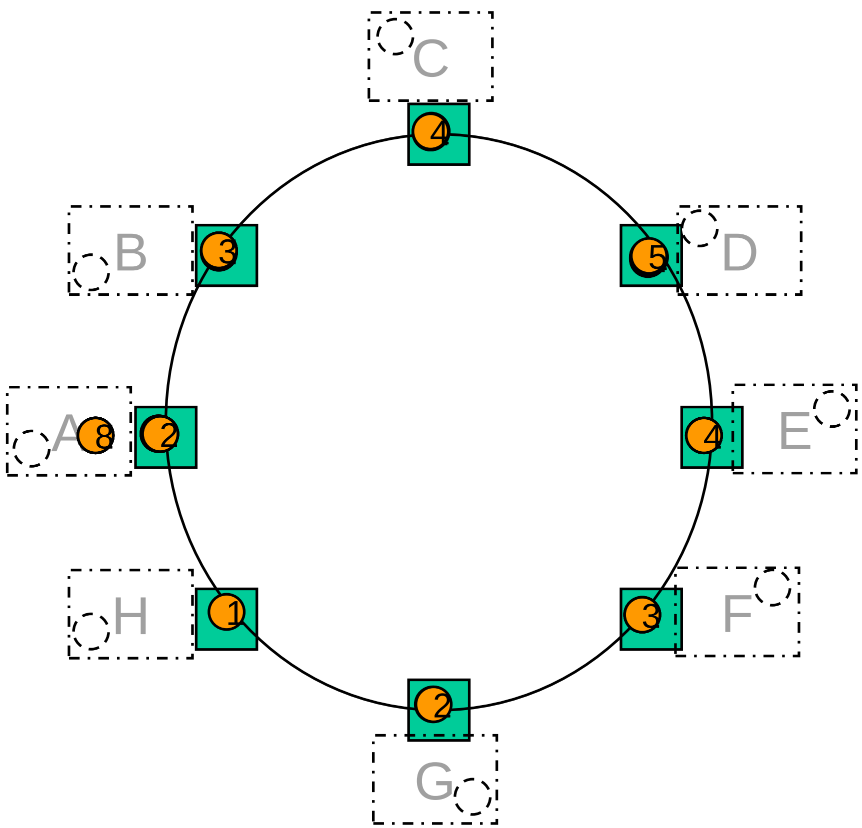


- Dutch one (DaCy)



The data cyclotron

- Construct a large main-memory ring buffer…
- A data chunk is loaded by a node into the ring…
- It continuously hops from node to node...
- Queries can attach at any node …



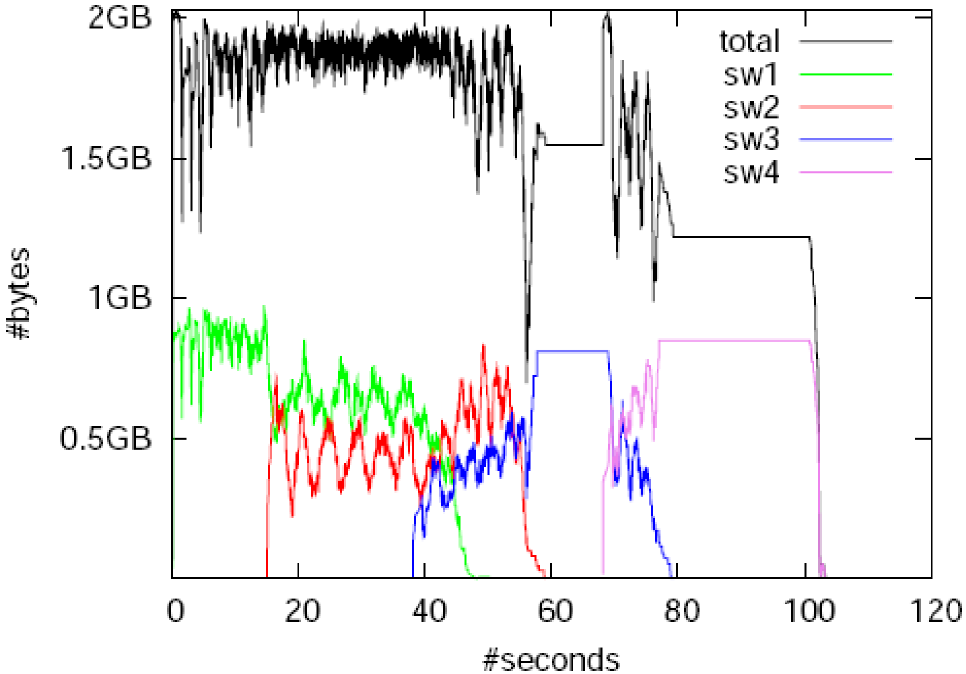
Hot-set management.

- Once the chunk stops being used by the queries, it is removed from the ring.
- In case you need to load new chunks, the less used ones are removed.
- LOI (Level Of Interest).

$$\mathrm{CAVG} = \frac{copies}{hops}$$

$$newLOI = \frac{LOI + CAVG \times cycles}{cycles}$$

Skewed Workload.

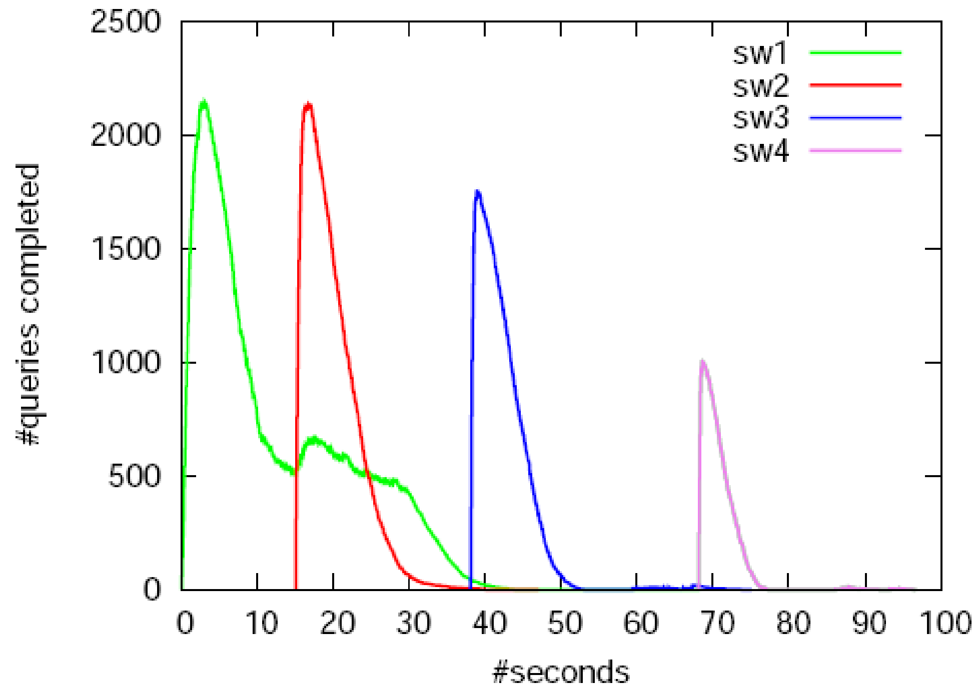


(a) Ring Load

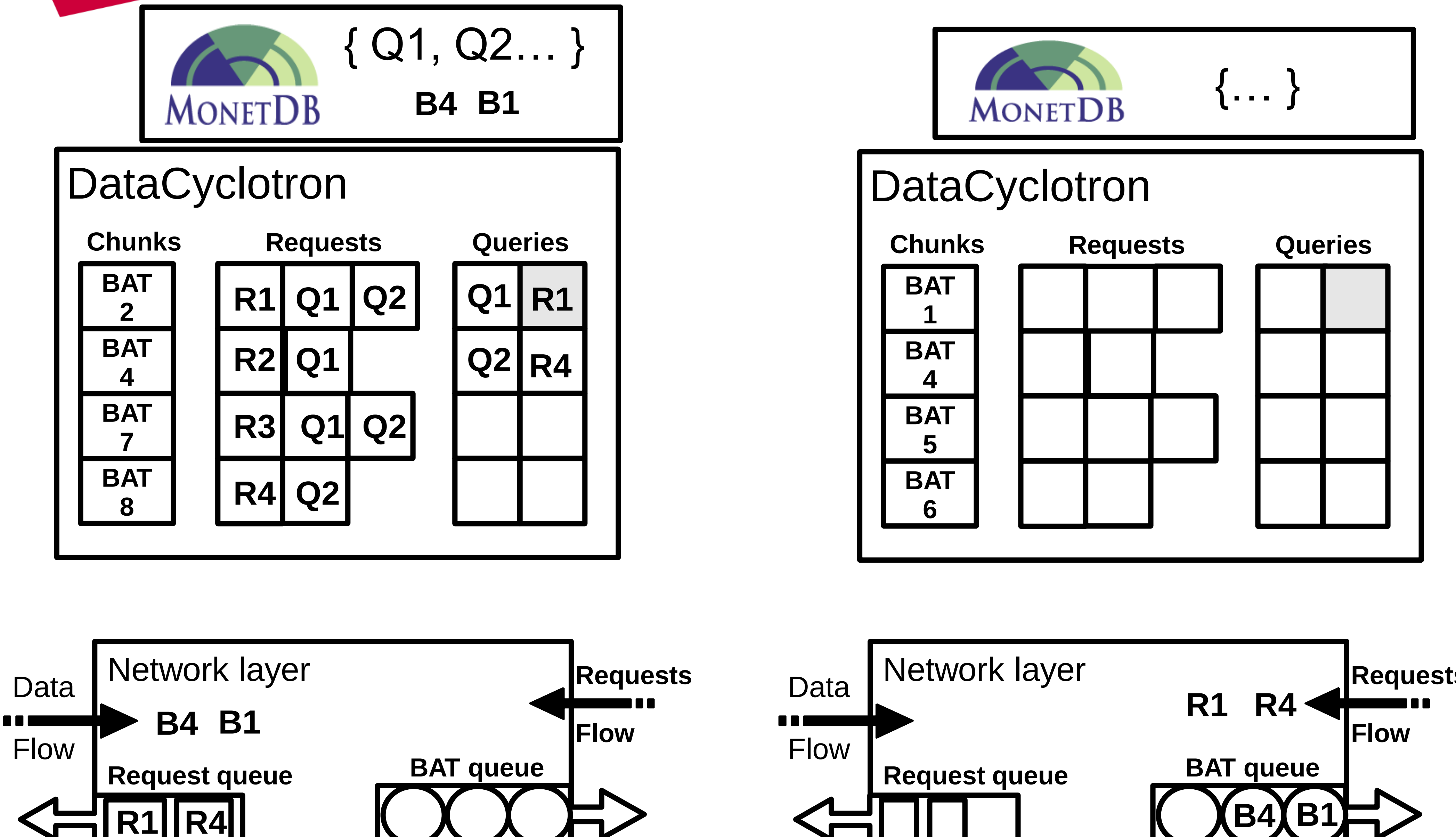
workload	SW1	SW2	SW3	SW4
skewed	3	5	7	9
start(sec)	0	15	37.5	67.5
end(sec)	30	45	67.5	97.5
queries/sec	200	300	400	500

TABLE IV

WORKLOAD DETAILS



(b) Query Throughput


MonetDB
{ Q1, Q2… }
B4 B1
DataCyclotron
Chunks
BAT 2
BAT 4
BAT 7
BAT 8
Requests
R1 Q1 Q2
R2 Q1
R3 Q1 Q2
R4 Q2
Queries
Q1 R1
Q2 R4
Network layer
Data Flow
B4 B1
Requests Flow
Request queue
R1 R4
BAT queue
MonetDB
{… }
DataCyclotron
Chunks
BAT 1
BAT 4
BAT 5
BAT 6
Requests
Queries
Network layer
Data Flow
R1 R4
Requests Flow
Request queue
BAT queue
B4 B1

DBMS integration.

- MonetDB

  select c.t_id from tab t, col c where c.t_id = t.id;

```
function user.s1_2():void;
   X1 := sql.bind("sys","tab","id",0);
   X6 := sql.bind("sys","col","t_id",0);
   X9 := bat.reverse(X6);
   X10 := algebra.join(X1, X9);
   X13 := algebra.markT(X10,0@0);
   X14 := bat.reverse(X13);
   X15 := algebra.join(X14, X1);
   X16 := sql.resultSet(1,1,X15);
   sql.rsCol(X16,"sys.c","t_id","int",32,0,X15);
   X22 := io.stdout();
   sql.exportResult(X22,X16);
end s1_2;
```

TABLE I

SELECTION OVER TWO TABLES

```
function user.s1_2():void;
   X2 := datacyclotron.request("sys","tab","id",0);
   X3 := datacyclotron.request("sys","col","t_id",0);
   X6 := datacyclotron.pin(X3);
   X9 := bat.reverse(X6);
   X1 := datacyclotron.pin(X2);
   X10 := algebra.join(X1, X9);
   X13 := algebra.markT(X10,0@0);
   X14 := bat.reverse(X13);
   X15 := algebra.join(X14, X1);
   X16 := sql.resultSet(1,1,X15);
   sql.rsCol(X16,"sys.c","t_id","int",32,0,X15);
   X22 := io.stdout();
   sql.exportResult(X22,X16);
   datacyclotron.unpin(X6);
   datacyclotron.unpin(X1);
end s1_2;
```

TABLE II

MAL PLAN AFTER DCOPTIMIZER

Summary

– We rotate the data through a ring of nodes using modern network technology, RDMA.

– A full fledged DBMS on each node.

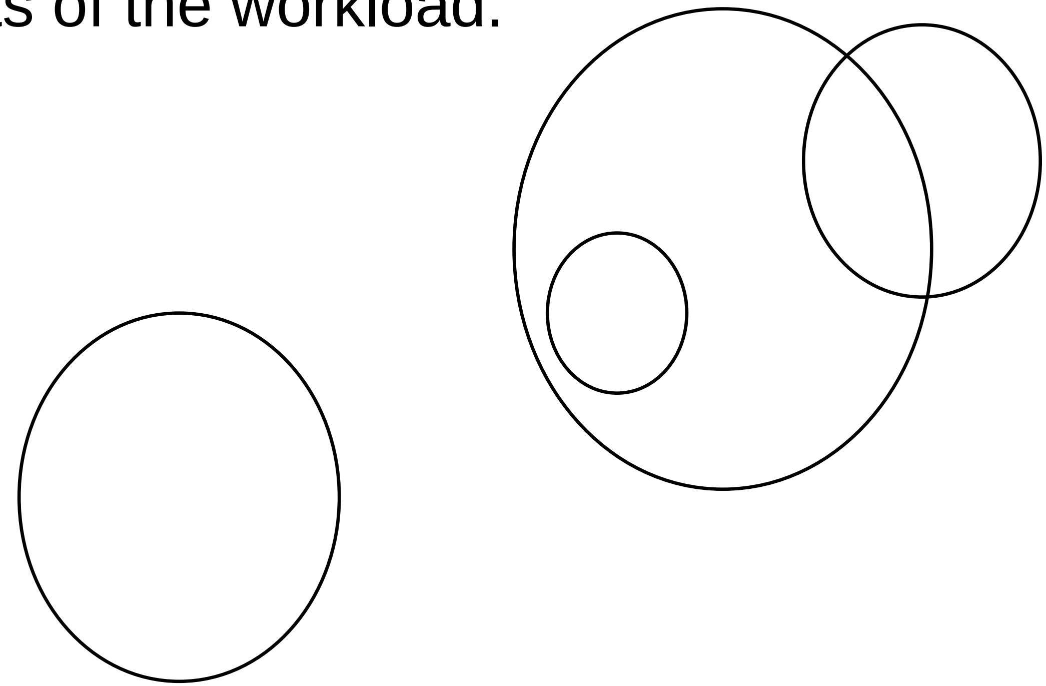
– Simple and efficient protocols to define the Hot data-set for skewed workloads...

– TPCH sf-100 runs on a 15-node ring, working towards scaling up to a 400-node ring

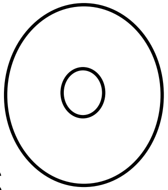
Old ideas become powerful on today's hardware

Future work

- Pulsating rings.
  - The ring grow and shrink to dynamically adapt to the requirements of the workload.



- Data Cyclotron Mesh.
  - Several overlapping pulsating rings.



Questions...
and
Remarks....

	Mid-1980s	**2009**	**Improvement**
Disk capacity	30 MB	500 GB	16667x
Maximum transfer rate	2 MB/s	100 MB/s	50x
Latency (seek + rotate)	20 ms	10 ms	2x
Capacity/bandwidth (large blocks)	15 s	5000 s	333x *worse*
Capacity/bandwidth (1KB blocks)	600 s	58 days	8333x *worse*
Jim Gray's Rule [11] (1KB blocks)	5 min.	30 hours	360x *worse*