CROSS-IDENTIFICATION AT THE CORE OF ASTRONOMY

6/16/2010

Tamás Budavári / The Johns Hopkins University

One Sky: SkyQuery

- Position – common across surveys
- Catalog federation
  - Static? How we store it?
    - SDSS-USNOB associations in SkyServer
  - Dynamic? Needs to be fast!
    - SkyQuery (2002), Open SkyQuery (2006)
- Next SkyQuery in the works
  - Zone algorithm: SDSS-GALEX in 20 min on 1 machine
  - Various selection of criteria including Bayesian

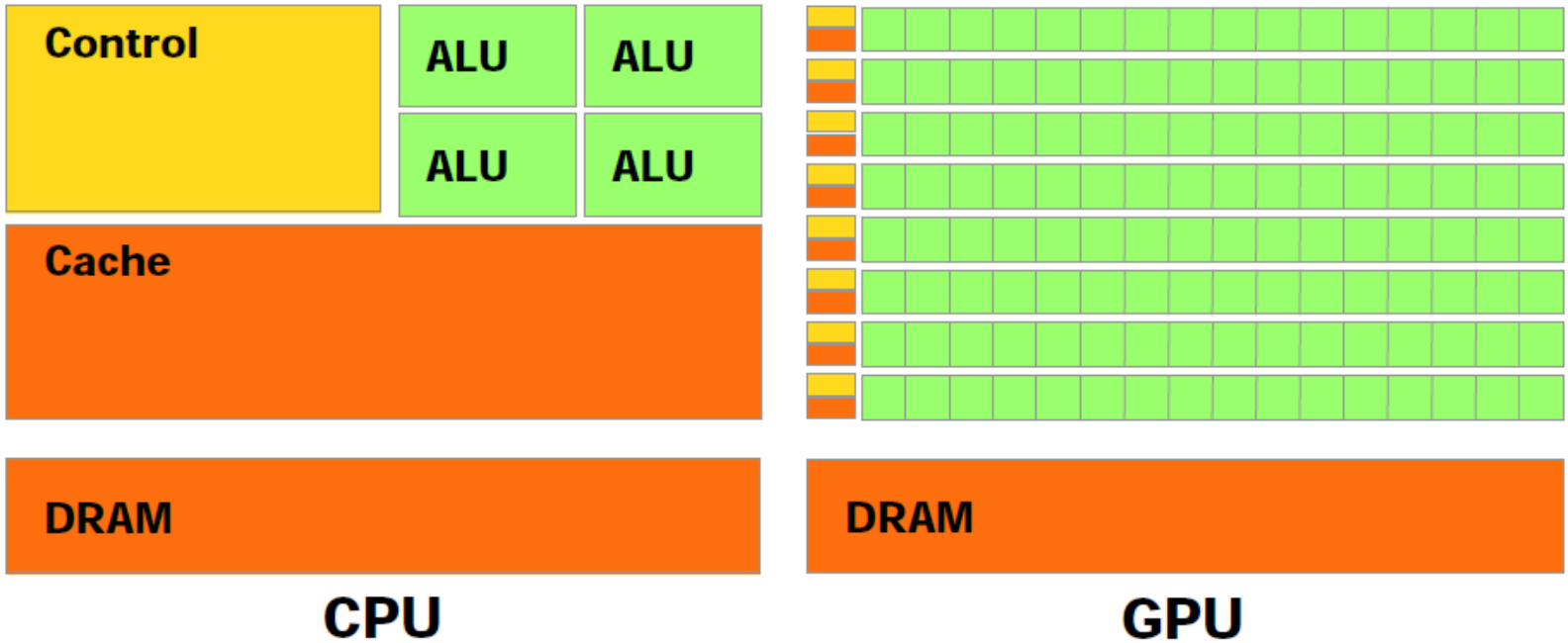
One Sky: SkyQuery

- Position – common across surveys
- Catalog federation
  - Static? How we store it?
    - SDSS-USNOB associations in SkyServer
  - Dynamic? Needs to be fast!
    - SkyQuery (2002), Open SkyQuery (2006)
- Next SkyQuery in the works
  - Zone algorithm: SDSS-GALEX in 20 min on 1 machine
  - Various selection of criteria including Bayesian

CPU Limited

New Moore's Law

- In the number of cores
- Faster than ever (for now)



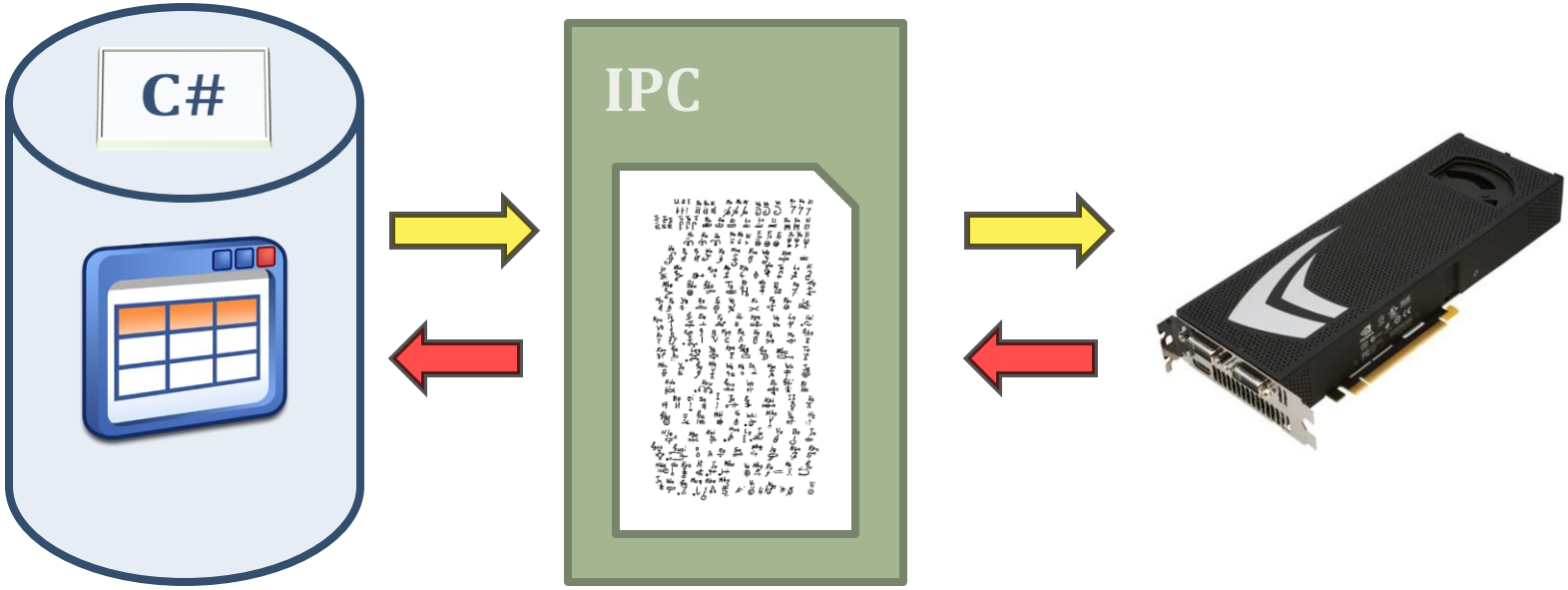
New Programming Paradigm

- 100s of cores – 27k parallel threads per GPU
  - Running a billion threads a second
- Forget the fancy old algorithms
  - Built on wrong assumptions
- Today CPU is free, RAM is slow
  - GPU has >50GB/s bandwidth
  - Still difficult to occupy the cores

Extending SQL Server

Tamás Budavári

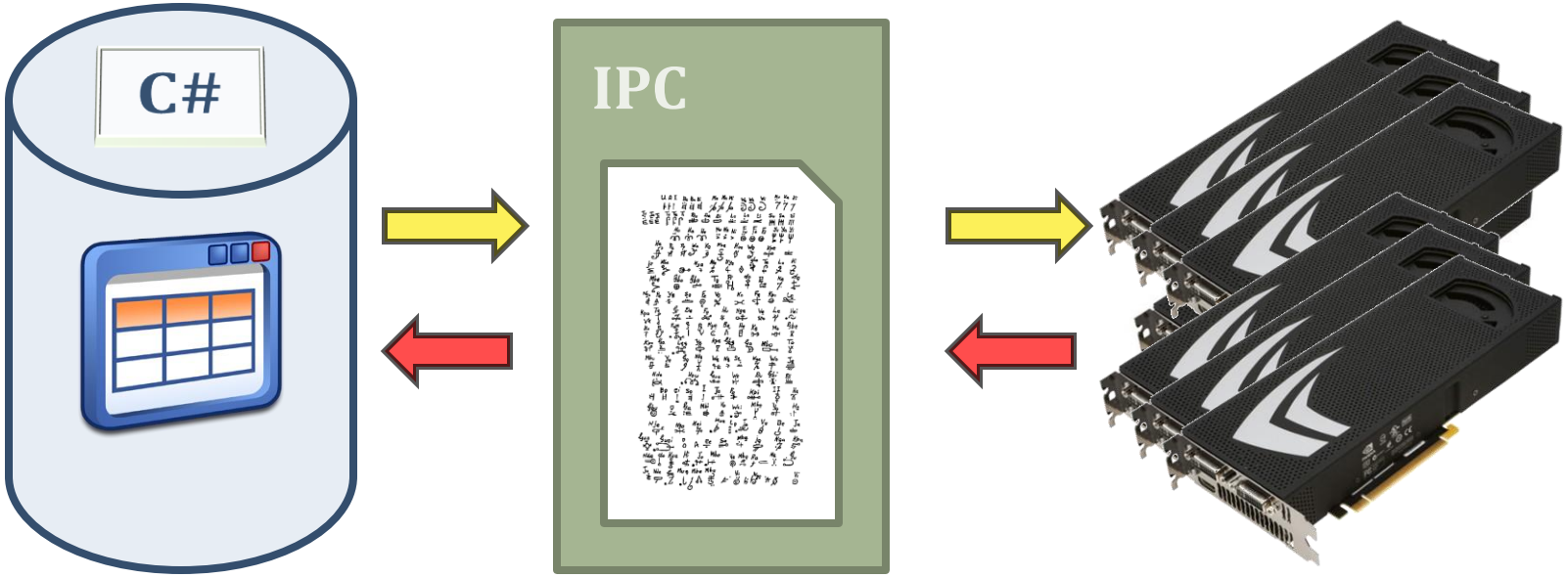
- Dedicated service for direct access
  - Shared memory IPC w/ on-the-fly data transform



Extending SQL Server

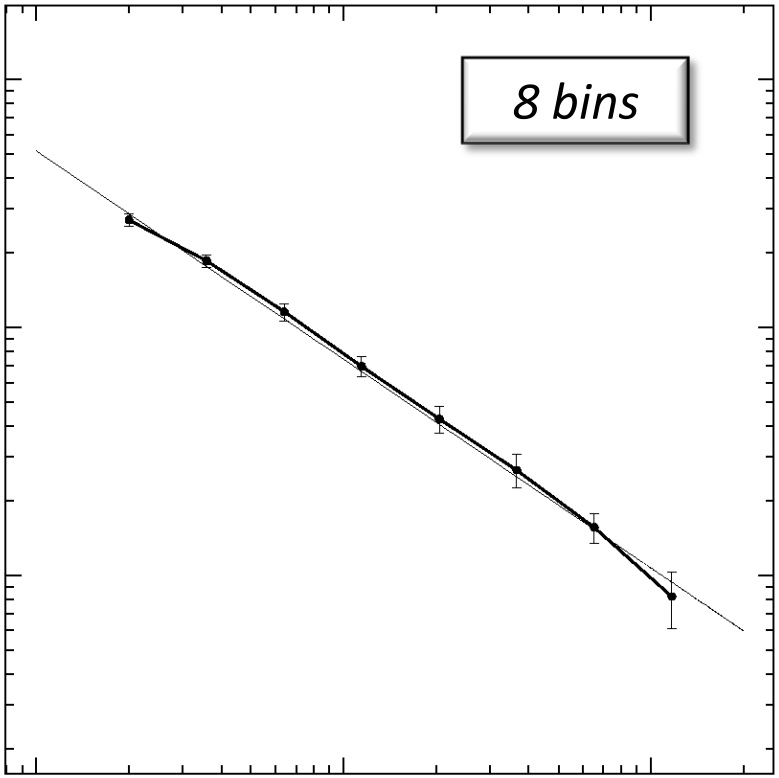
Tamás Budavári

- Dedicated service for direct access
  - Shared memory IPC w/ on-the-fly data transform

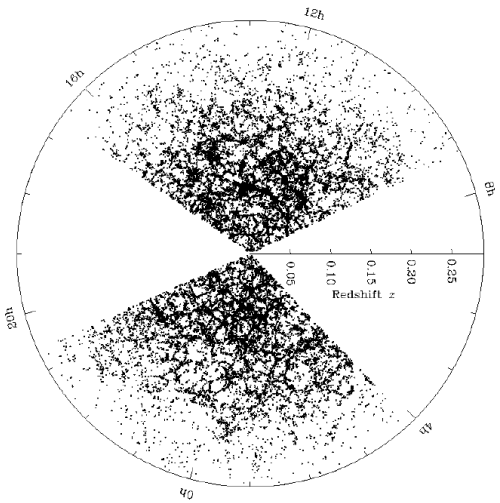


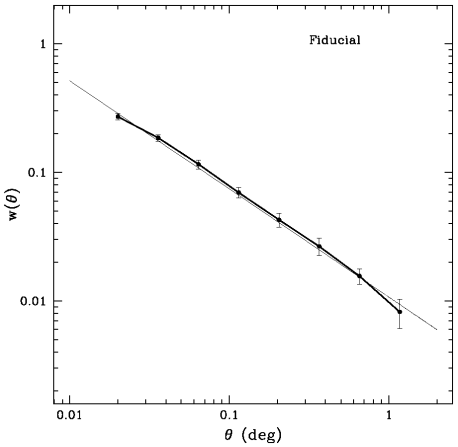
An Example

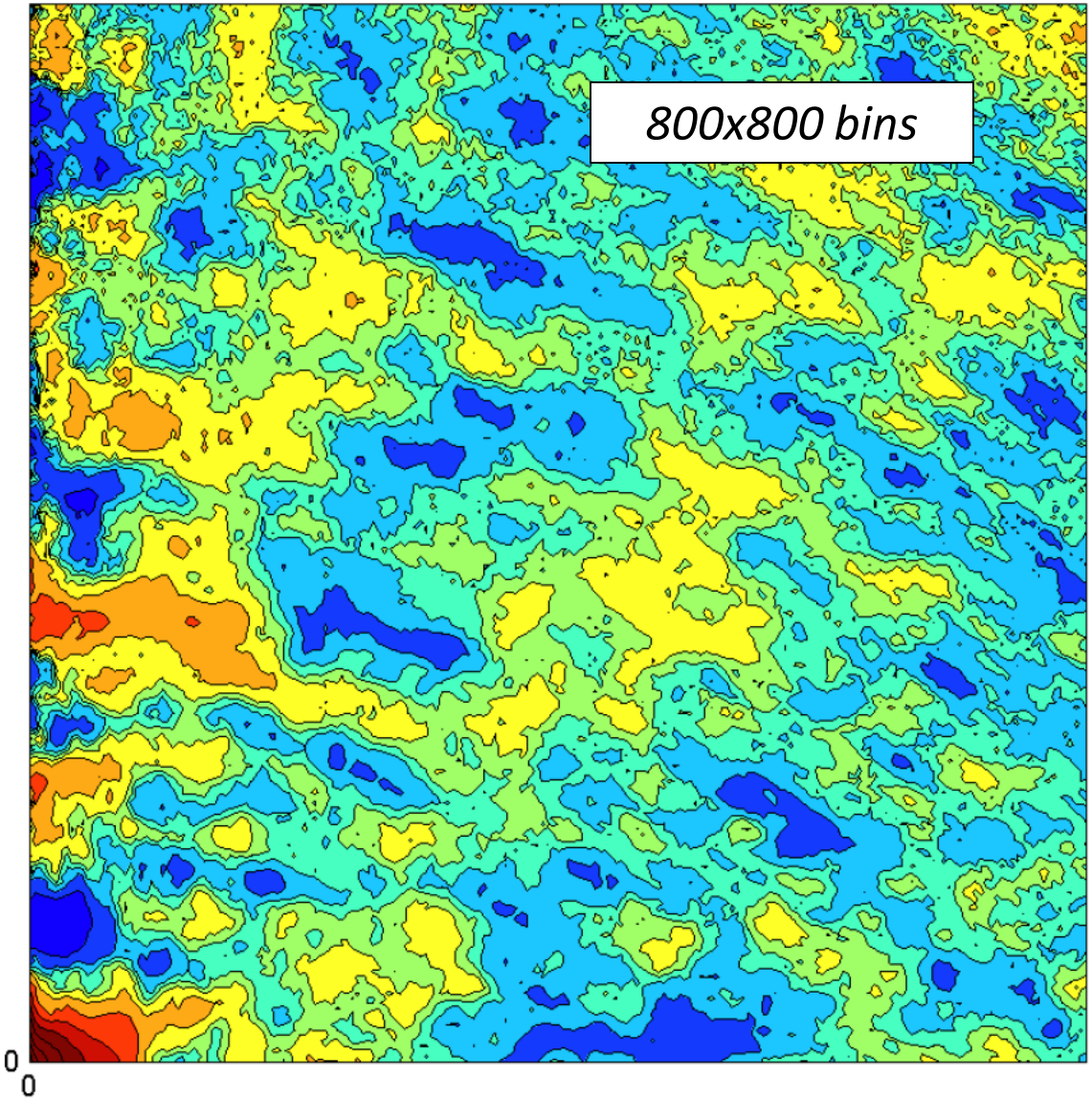
- Correlation functions
  - From pair-counts
- State of the art
  - Dual-tree traversal
- High resolution bins?
  - Just like brute force



Sloan DR7


0h
20h
40h
12h
16h
8h
0.05
0.10
0.15
0.20
0.25
Redshift z


Fiducial
1
0.1
0.01
w(θ)
0.01
0.1
1
θ (deg)


800x800 bins
0
0

All Done Inside the Database

- Pair counts computed on GPU
  - Returns 2D histogram as a table (i, j, cts)
- Calculate the correlation fn in SQL

```
select dd.i, dd.j,  dd.cts as dd, dr.cts as dr, rr.cts as rr
    (@Nrr*CONVERT(float,dd.cts)/@Ndd - 2*@Nrr*CONVERT(float,dr.cts)/@Ndr + CONVERT(float,rr.cts))
        / CONVERT(float,rr.cts) as xi
from     dbo.PairCounts(@maxmpc, @nbin, @qryD, @nD, null) dd
    join dbo.PairCounts(@maxmpc, @nbin, @qryR, @nR, null) rr  on dd.i = rr.i and dd.j = rr.j
    join dbo.PairCounts(@maxmpc, @nbin, @qryDR, @nD, @nR) dr  on dd.i = dr.i and dd.j = dr.j
go
```

- Can also do async parallel GPU jobs

Summary?

- Statistical challenges!
  - Bayesian cross-identification
- Which architecture?
  - CPUs w/ 6 cores
  - GPUs w/ 480 cores
  - Cloud computing?
- New design and programming paradigm!