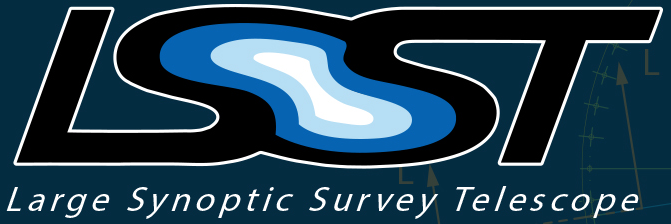
Qserv: a distributed shared-nothing database for the LSST catalog

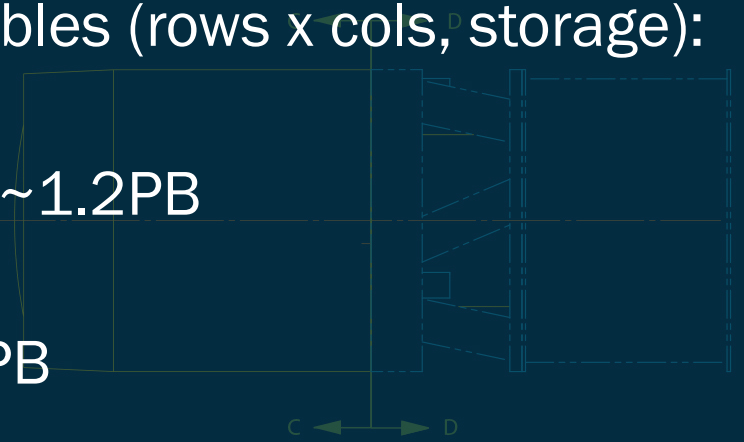
Fritz Mueller
SLAC National Accelerator Laboratory



The LSST L2 Catalog

- Data, by DR11:
  - ~60T rows (mostly ForcedSource)
  - ~10PB (mostly Source + ForcedSource + Object extra)

- Breakdown of most significant tables (rows x cols, storage):
  - Object: ~47B x 330, ~100TB
  - Object extra: ~1.5T x 7,600, ~1.2PB
  - Source: ~9T x 50, ~5PB
  - ForcedSource: ~50T x 6, ~2PB

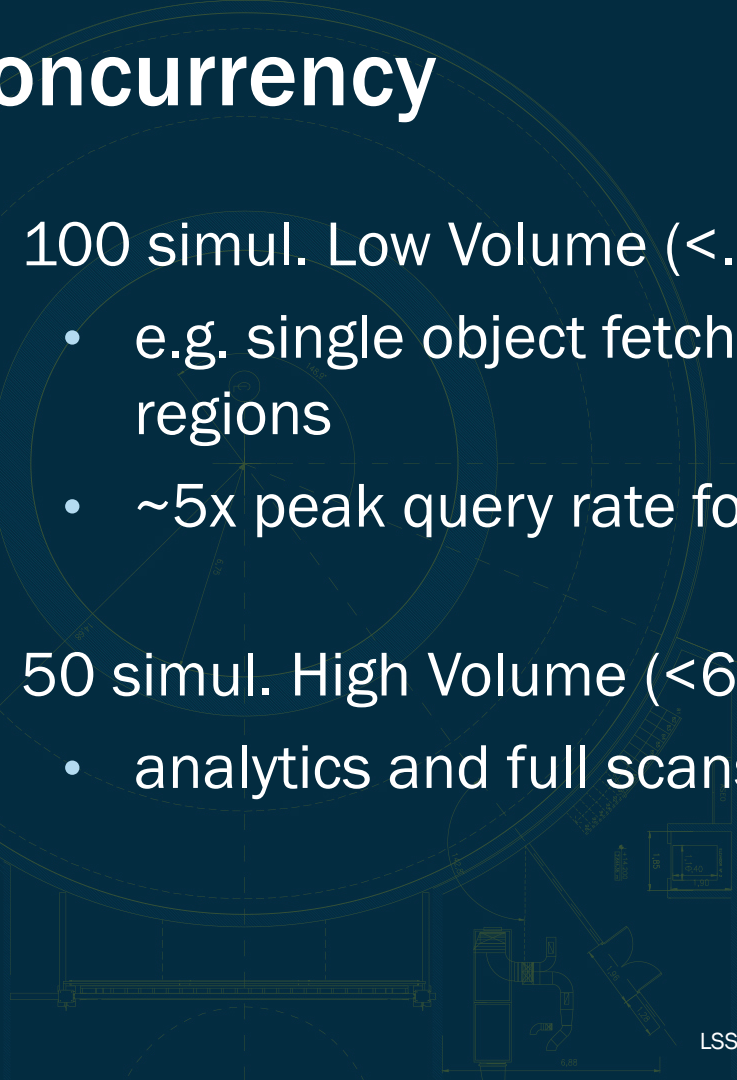


Analytics

- In a region
  - Get an object or data for small area - <10 sec
- Across entire sky
  - Scan through billions of objects - ~1 hour
  - Deeper analysis (Object_*) - ~8 hours
- Analysis of objects close to other objects
  - ~1 hour, even if full-sky
- Analysis that requires special grouping
  - ~1 hour, even if full sky
- Time series analysis
  - Source, ForcedSource scans - ~12 hours
- Cross match & anti-cross match with external catalogs
  - ~1 hour

Concurrency

- 100 simul. Low Volume (<.5GB) @ 10/sec
  - e.g. single object fetch or small (10's of arcmin) spatial regions
  - ~5x peak query rate for SDSS SkyServer
- 50 simul. High Volume (<6GB) @ 20/hr
  - analytics and full scans



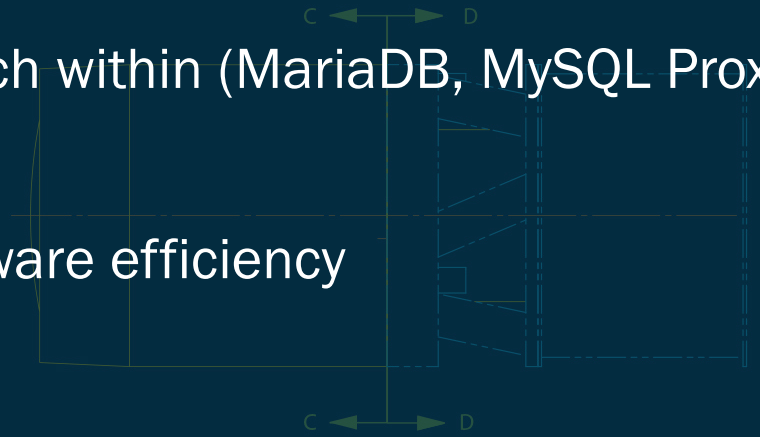
The Qserv Approach

- Shared-nothing MPP RDBMS (throughput, horizontal scaling)
- Spatial partitioning with overlap (near-neighbor self-joins)
- Shared scans (concurrent query load)
- Replicated data (resiliency)
- Fixed-purpose, dedicated hardware (cost, predictability)

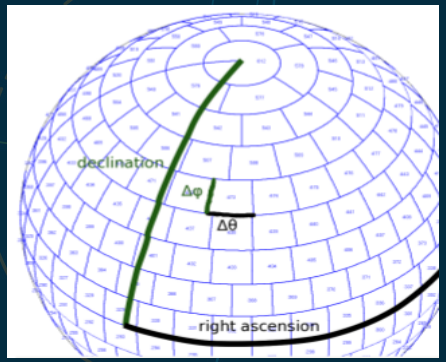
Build it ourselves, leverage existing tech within (MariaDB, MySQL Proxy, XRootD, Google protobuf, Flask)

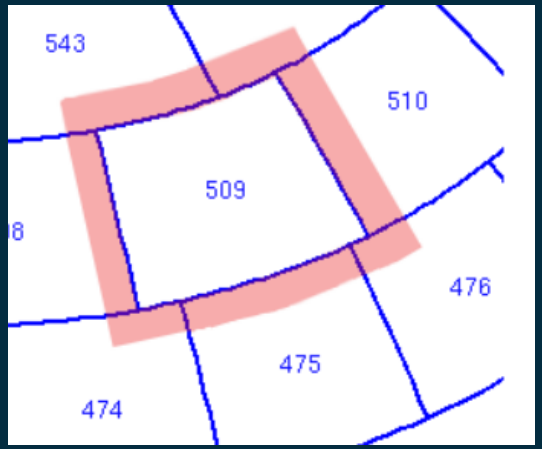
Design optimized for use case + hardware efficiency

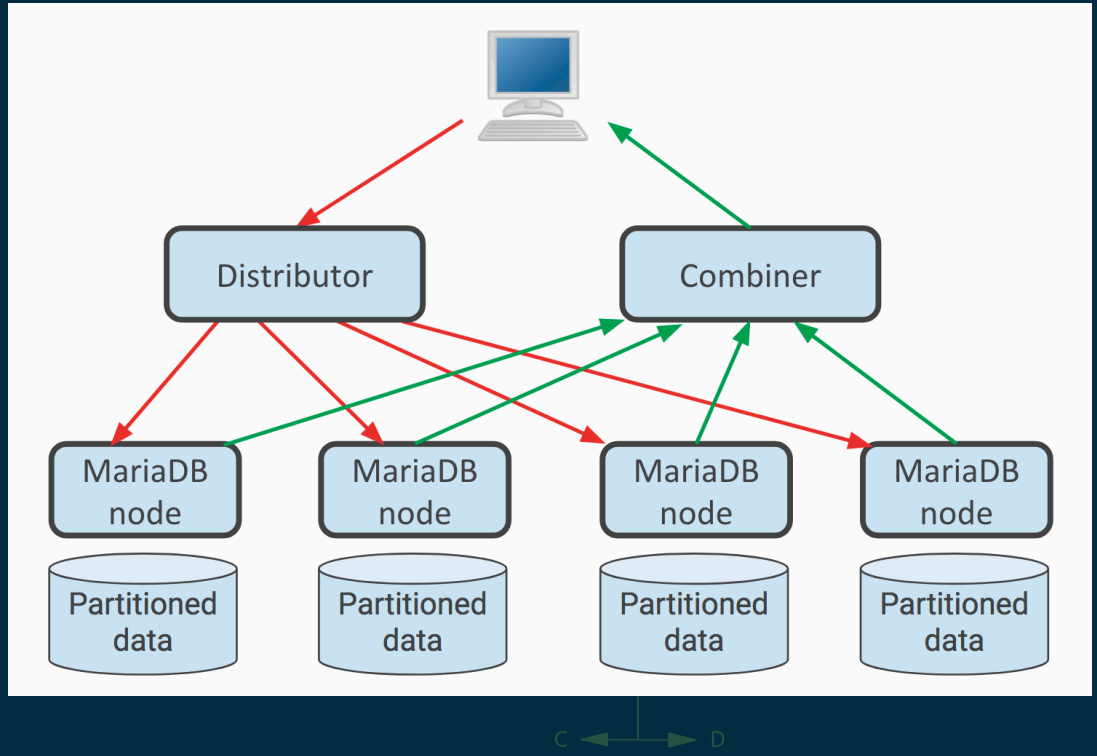
100% open source



Shared-nothing MPP







Spherical Partitioning

Robust spherical geometry in the database

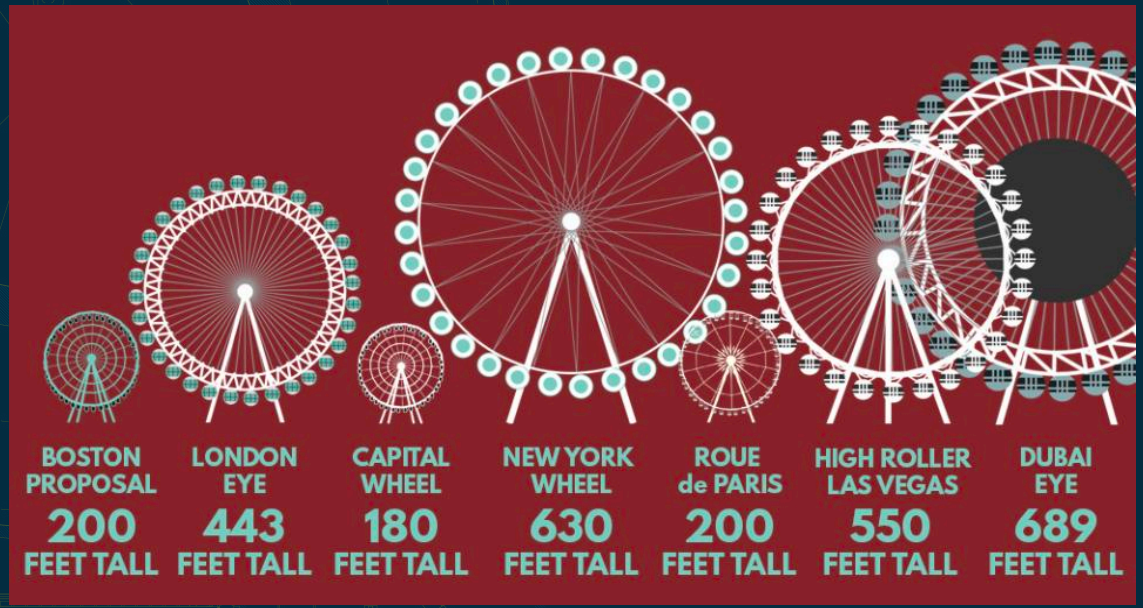
- 0/360 RA wrap around, well behaved poles, convex polygons, accurate distance computation, angular distance
- Point-in-spherical-region tests (circle, ellipse, box, convex polygon)
- Custom (HTM-based) UDFs (https://github.com/smonkewitz/scisql)

Optimized spatial joins for neighbor queries, cross-match

- Spherical partitioning with overlap
- Director table, secondary index
- Two-level, 2nd level materialized on-the-fly

Shared Scans

- Continuous, sequential scans through data, including L3 distributed tables
- (Non-interactive) queries attached to appropriate running scan



Interaction: Spatial Restriction

```
qserv_areaspec_box(lonMin, latMin, lonMax, latMax)

qserv_areaspec_circle(lon, lat, radius)

qserv_areaspec_ellipse(semiMajorAxisAngle,
semiMinorAxisAngle, posAngle)

qserv_areaspec_poly(v1Lon, v1Lat, v2Lon, v2Lat, …)
```

```
SELECT objectId FROM Object
WHERE qserv_areaspec_box(2,89,3,90) AND …
```

Interaction: Common Query Types

`SELECT … FROM Object`
- massively parallel

`SELECT … FROM Object WHERE qserv_areaspec_box(…)`
- selection inside chunks that cover requested area, in parallel

`SELECT … FROM Object JOIN SOURCE USING (objectId)`
- massively parallel without any cross-node communication

`SELECT … FROM Object WHERE objectId = <id>`
- quick selection inside one chunk

Example queries: http://ls.st/ed4

Interaction: Query Limitations

Only a SQL subset is supported. For example:

- Spatial constraints (must use User Defined Functions, must appear at the beginning of WHERE, only one spatial constraint per query, arguments must be simple literals, OR not allowed after area qserv_areaspec_*)
- Expressions/functions in ORDER BY clauses are not allowed
- Sub-queries are NOT supported
- Commands that modify tables are disallowed
- MySQL-specific syntax and variables not supported
- Repeated column names through * not supported

Deployments

Production Target:

- ~500 nodes in 2 international data-centers

Development cluster (CC-IN2P3):

- 400 cores, 800 GB memory, 500 TB storage
- ~70 TB synthetic dataset on 2 x 25 nodes
- ~100 TB synthetic database coming up

Prototype Data Access Center (NCSA):

- 500 cores, 4 TB memory, 700 TB storage
- ~25 TB science dataset (SDSS Stripe 82 + WISE) on 30 nodes
- ~100 TB science dataset coming up (+ WISE n-band + HSC reprocessing)

Recent scale tests: http://ls.st/ucx

Active Work

- Data distribution/replication
- Resource management (user quotas, query estimation)
- User datasets (mydb)
- Next-to-data processing
- Deployment and operation improvements
- Data ingest tooling

Want to learn more? Many more details at http://ls.st/LDM-135