Gamma

Anastassia Ailamaki
http://www.cs.cmu.edu/~natassa

History
- DIRECT 1977 - 84
  - early database machine project
  - showed parallelism useful for db apps
- Flaws curtailed scalability
  - shared memory
  - central control of execution
- Gamma 1984 - 92

Key Ideas
- Shared-nothing
- Hash-based parallel algorithms
- Horizontal partitioning
Gamma Hardware (v2.0)
- iPSC/2 Intel hypercube
- 32 x386 processors
- 8MB of memory
- 330MB Maxtor drive / node (45KB cache)
- Routing modules
  - 2.8 Mb/s
  - full duplex, serial, reliable

Gamma Software
- OS: NOSE
- multiple, lightweight processes with shared memory
- Entire DB in one NX/2 process
- Details: renaming nodes
- 10% CPU used for copying
- Excessive interrupts during I/O

Storage Organization
- Horizontal partitioning (user selectable)
  - round-robin
  - hashed
  - range partitioned
- Clustered index on different attributes
- Partition relations should have based on ‘heat’
Gamma Process Structure

- Catalog manager
  - repository for db schema
- Query manager
  - one associated with each user
- Scheduler processes
  - coordinates multi-site queries
- Operator processes
  - each executes single relational operator

Query Processing

- ad-hoc and embedded query interfaces
- standard parsing, optimization, code gen.
- left deep trees only
- hash joins only
- at most two join operators active simultaneously
- split tables (partitioning, joining)
Split Table
Directs operator output to the appropriate node

Parallel Hash Join

Scheduling Operators
Selections and Joins

- Selections
  - Start selection operator on each node
  - Exclusion of nodes for hash and range partitioning
  - Throughput considerations
    - One page read-ahead

- Joins
  - Partition into buckets, join buckets
  - Implemented: sort-merge, Grace, Simple, Hybrid
  - Parallel Hybrid

More operations

- Aggregation
  - Compute partial results for each partition
  - Hash on "group-by" attribute

- Updates
  - Standard techniques
  - How update partitioning attribute?

Concurrency Control

- 2PL
- Granularity: file and page
- Modes: S, X, IS, IX, SIX
- Local lock manager and deadlock detector
- Wait-for graph
- Centralized multi-site lock detector
- Period doubles/halves on deadlock/deadlock
Failures and Recovery

- Log records with LSNs
  - LSN = f (node number, local sequence number)
- Processor i directs log records at log manager
  - (i mod M), where M = # log mgrs.
- Standard WAL protocol
  - local log manager keeps T clean/unfixed pages
  - reduces time to wait for log to be written
- ARIES

Node Failure

- Availability in spite of processor or disk fail
- Mirrored disk (Tandem)
- Interleaved declustering (Teradata)
- Chained declustering
- Load redirection results in \(1/n\) increase

Fault Tolerance

- Primary copy and backup copy
- Objective: availability + easy balance

- Chained Declustering
  - Primary: \(i \mod M\)
  - Backup: \((i+1) \mod M\)
- Interleaved declustering
  - Divide fragments into \(N-1\) parts
  - Store parts in all disks but the one containing primary