Adaptive Transaction Processing on Hardware Islands

Danica Porobic, Erietta Liarou, Pinar Tözün, Anastasia Ailamaki
scaling up OLTP on multisockets

Throughput

# of sockets

1 2 4 8
scaling up OLTP on multisockets

Throughput

1 2 4 8

# of sockets
scaling up OLTP on multisockets

Multisocket servers are severely underutilized
multisocket multicores

Communication latencies vary by an order of magnitude
multisocket multicores

Communication latencies vary by an order of magnitude
processor trends

2005

ILP
pipelining
multithreading

multisocket multicores
(CMP)
processor trends

2005

ILP
pipelining
multithreading

2020

multisocket multicores (CMP)
heterogeneous CMP
OLTP on Hardware Islands
OLTP on Hardware Islands

Shared-everything

DB

data indexes  Logs
OLTP on Hardware Islands

**Shared-everything**

`schematic diagram`

**Shared-nothing**

`schematic diagram`
OLTP on Hardware Islands

Shared-everything

Island shared-nothing

Shared-nothing

data
indexes

Logs

DB

DB

DB

DB

DB
OLTP on Hardware Islands

Shared-everything

- stable
- no optimal

Island shared-nothing

- Robust middle ground

Shared-nothing

- Fast
- sensitive to workload

challenges

* optimal configuration depends on workload and hardware
* expensive repartitioning due to physical data movement
ATraPos: Adaptive Transaction Processing

- No unnecessary inter-socket synchronization
- Workload- & hardware-aware partitioning
- Lightweight monitoring & repartitioning

Hardware and workload-aware shared-everything adaptive system
Critical path of transaction execution

transaction to thread

System state

threads

Data

Many accesses to shared data structures
PLP: Physiological partitioning

System state is still shared

ATraPos: Island-aware SE

System state

threads

Core Core Core Core Core Core

System state

Core Core Core Core
Perfectly partitionable workload

- Shared-nothing
- ATraPos
- PLP
- Centralized shared-everything

Island awareness brings scalability

8 socket x 10 core
800K row dataset
Probing one row
Naive partitioning and placement

8 socket x 10 core
800K rows per table
Probing 1 row each from A and B

Probe A
Probe B

Cores are overloaded with contending threads
A TraPos partitioning and placement

Ignoring Islands brings synchronization overhead

8 socket x 10 core
800K rows per table

Probing 1 row each from A and B

Probe A

Probe B

Throughput (KTPS)

PLP
A TraPos HW-aware
A TraPos Load balanced

4.4x
ATraPos partitioning and placement

ATraPos: load balance and reduce synchronization

8 socket x 10 core
800K rows per table

Probing 1 row each from A and B

ATraPos HW-aware
ATraPos Load balanced
ATraPos

Throughput (KTPS)

4.8x

0
300
600
900
1200
1500
1800
2100

PLP

ATraPos

Probe A

Probe B
ATraPos monitoring and repartitioning

**Goal:**
- balance the load
- minimize synchronization
Repartitioning multi-rooted B-trees
Repartitioning multi-rooted B-trees

Splitting and merging B-trees access few pages
ATraPos repartitioning cost

Repartitioning cost (ms)

- merge (solid blue line)
- split (dashed blue line)
- rearrange (split+merge) (dotted blue line)

Number of repartitioning actions

8 socket x 10 core
800K row table
ATraPos repartitioning cost

Repartitioning takes < 200msec
TATP speedup over PLP

Normalized throughput

GetSubData  GetNewDest  UpdSubData  Mix

ATraPos

PLP

8 socket x 10 core
800K subscribers
TATP speedup over PLP

ATraPos improves performance by 3.1-6.7x
Adapting to workload skew

Throughput (MTPS) vs. Time (s)

- Monitoring
- Repartitioning

ATraPos

8 socket x 10 core
800K subscribers
TATP GetSubData
Adapting to workload skew

ATraPos detects skew and quickly adapts
Adapting to workload skew

ATraPos detects skew and quickly adapts

Throughput (MTPS) vs. Time (s)

- Static
- ATraPos

Monitoring
Repartitioning
8 socket x 10 core
800K subscribers
TATP GetSubData

50% requests to 20% data
Adapting to any workload type

Throughput (KTPS) vs. Time (s)

- Static
- ATraPos

Events:
- Monitoring
- Repartitioning
- GetNewDest
- Mix
- UpdSubData

System Specifications:
- 8 socket x 10 core
- 800K subscribers
- TATP
Adapting to any workload type

ATraPos gracefully adapts to any changes
Adaptive OLTP for Islands

challenges

* optimal configuration depends on workload and hardware
* expensive repartitioning due to physical data movement

ATraPos

* minimal intersocket access to the critical path
* workload & hardware-aware partitioning & placement
* lightweight monitoring and repartitioning

Thank you!