Don’t be a victim of your own success!

E-Store: Elastic Resource Provisioning for OLTP Databases

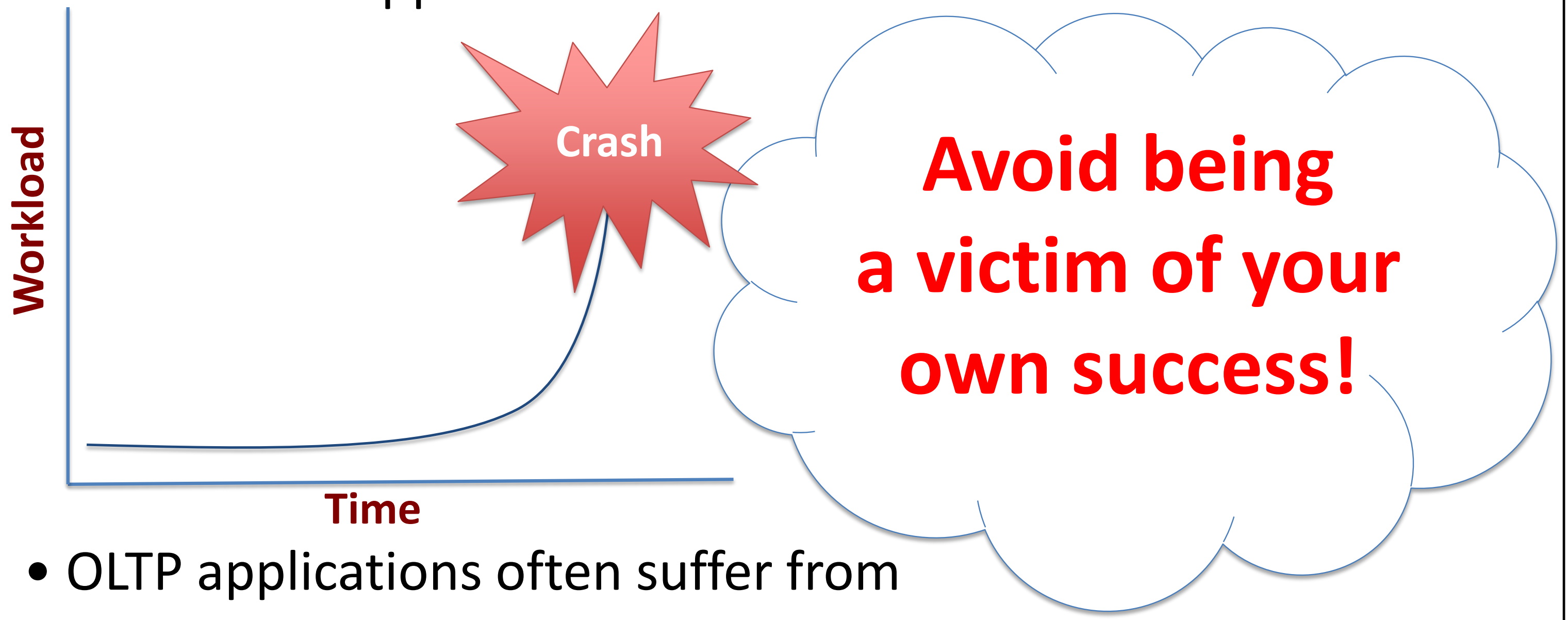
Rebecca Taft, Essam Mansour, Marco Serafini, Jennie Duggan, Aaron J. Elmore, Ashraf Aboulnaga, Andrew Pavlo, Michael Stonebraker

CSAIL

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Problem and Motivation

- Online Transaction Processing (OLTP) is the core of many business applications and web services



- OLTP applications often suffer from high load skew and variation

Extreme Skew

40% - 60% of the trading volume on the New York Stock Exchange (NYSE) occurs in 40 specific stocks

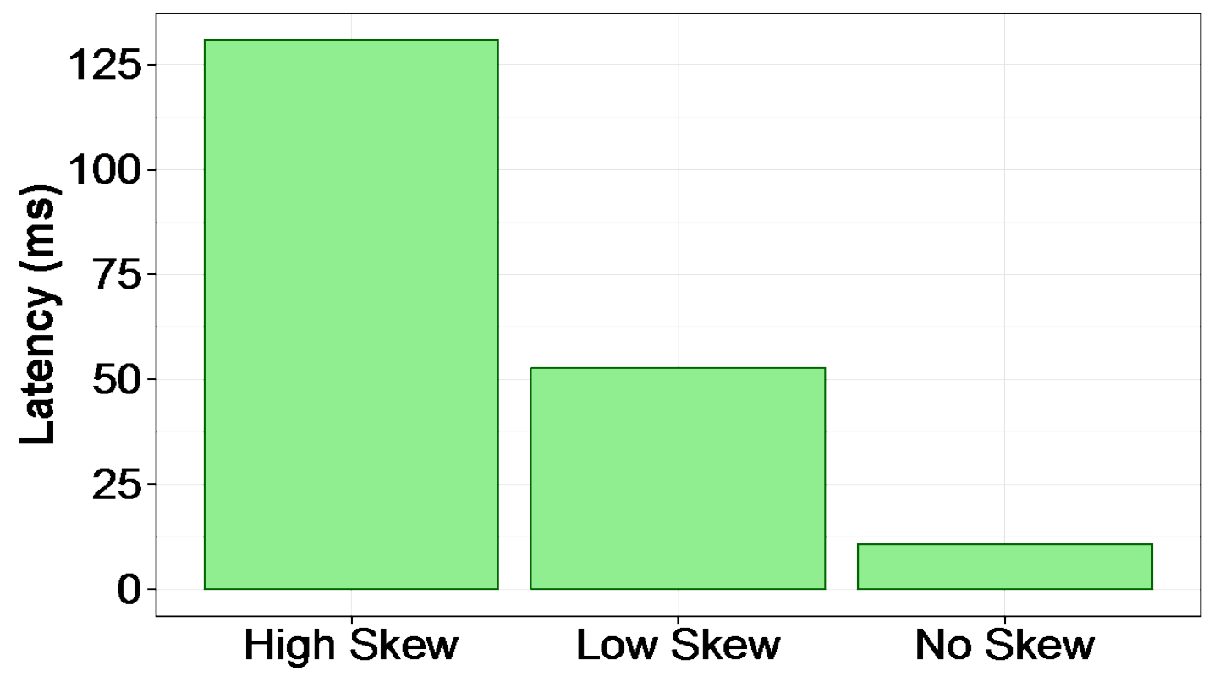
Time-Varying Skew

“follow the sun”: load moves around the globe following daylight hours

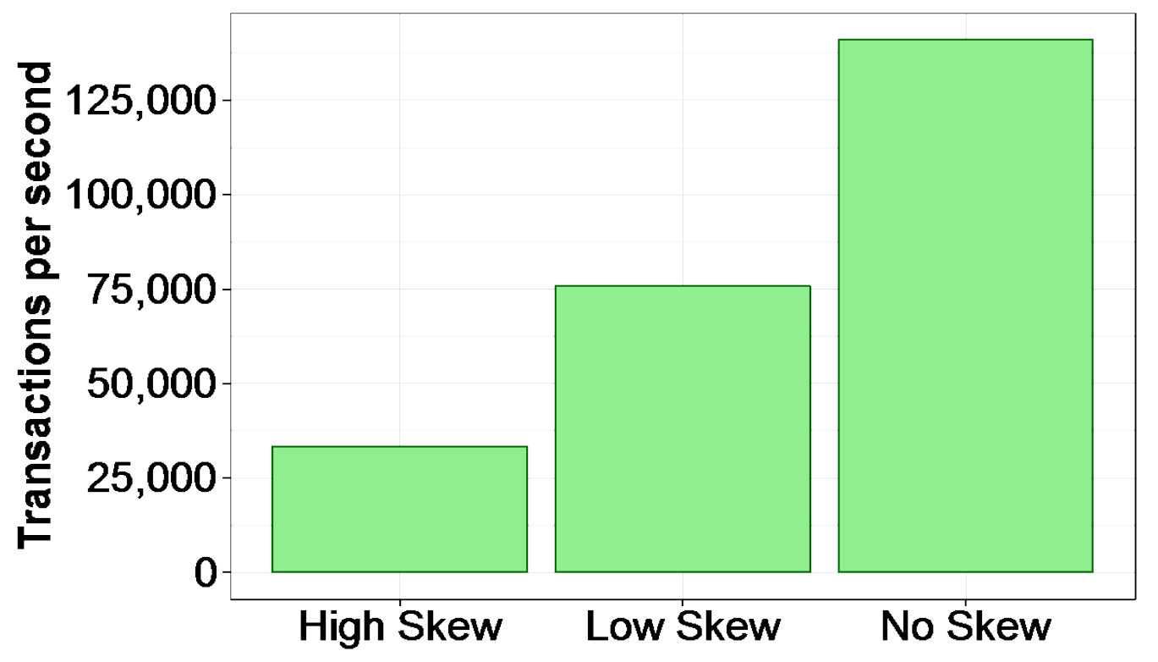
Load Spikes and Periodic Changes

The volume on the NYSE during the first and last ten minutes of the trading day is 10X higher than at other times. Seasonal travel companies (e.g., ski resorts) experience seasonal variation in load.

Under load skew or load variation, average latency will increase, perhaps by an order of magnitude, and average throughput will decrease dramatically



Average Latency (Milliseconds)



Average Throughput (Txns Per Second)

How to avoid performance degradation?

- Provision computing resources to support peak load (What is peak load? Very expensive)
- Limit the load on the system (Miss the window for success)

The goal for this work is to enable OLTP systems to elastically scale out and scale in computing resources to maintain the required performance regardless of load skew or load variation.

Our Solutions

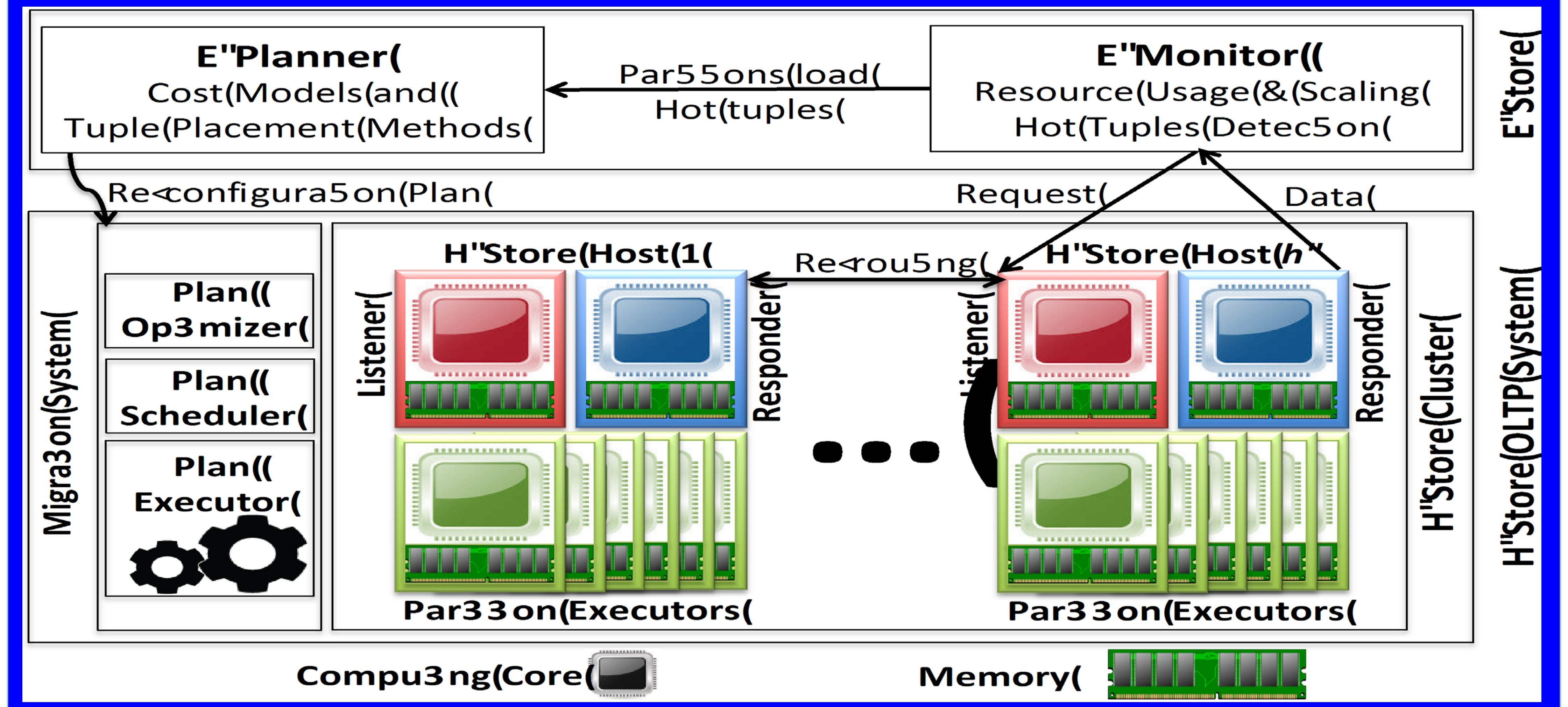
•Why is this hard?

•Maintaining the consistency of the database

•Reducing the data migration cost when scaling out or in

•Identifying fine-grained hotspots (tuples) that need extra resources

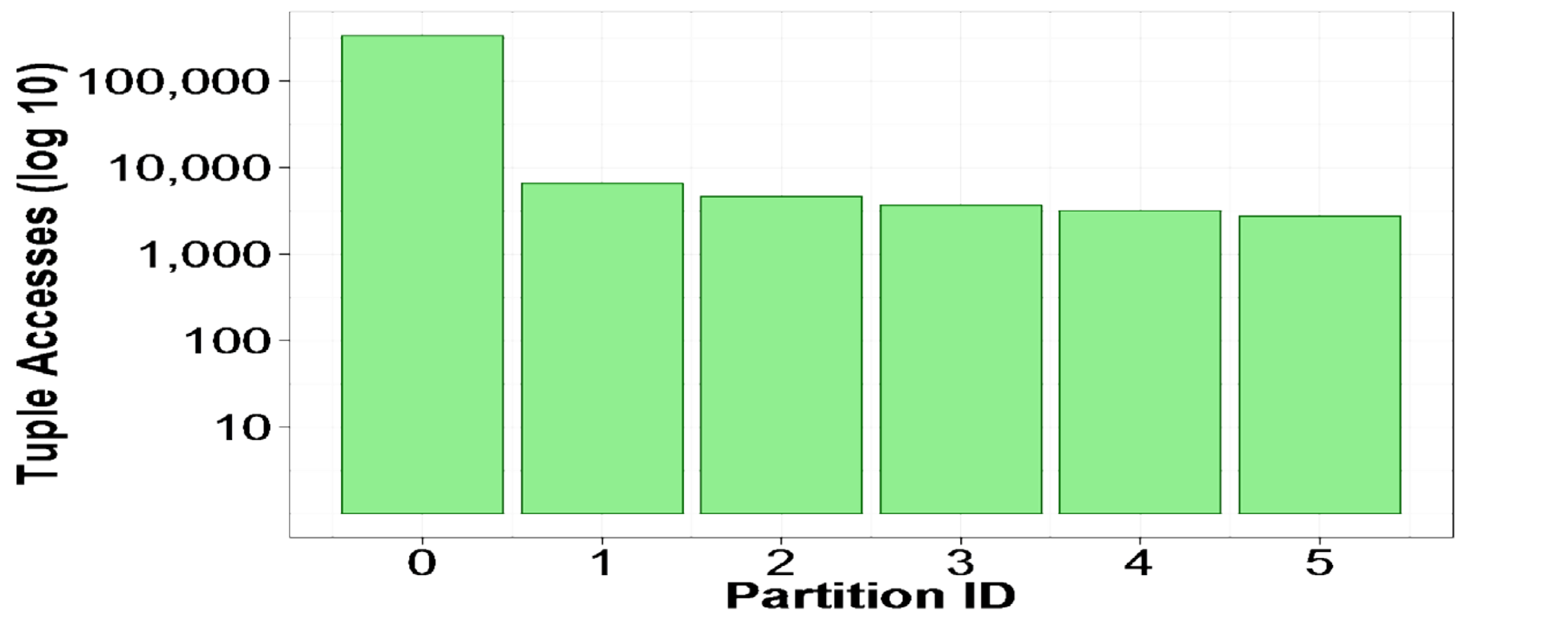
E-Store Architecture



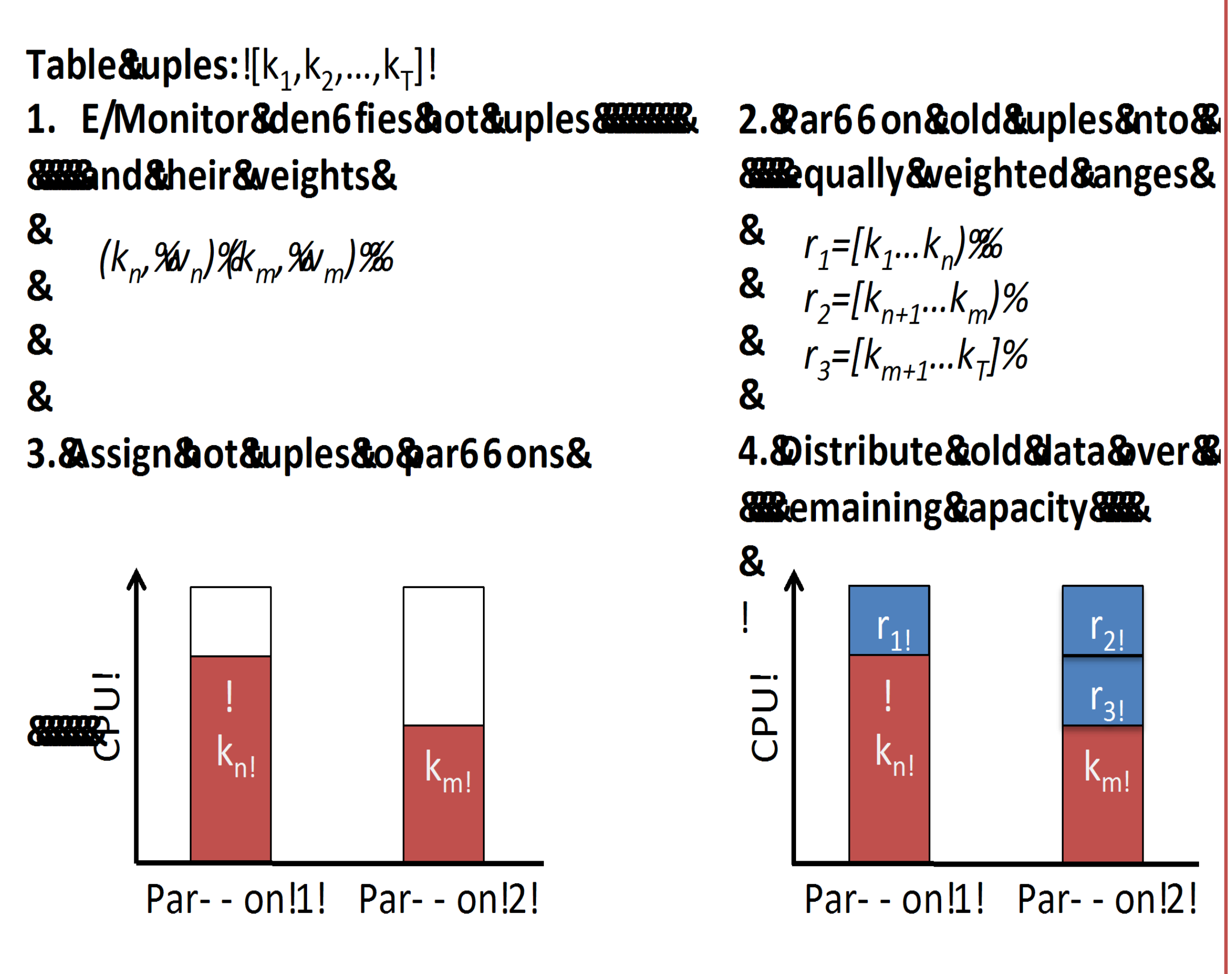
E-Store is a planning and reconfiguration framework for shared-nothing and partitioned OLTP DBMSs. Our main contribution is a comprehensive elasticity framework that can deal with load skew or load variations. The E-Store framework consists of two main components **E-Monitor** and **E-Planner**. To dynamically balance the access load among the partitions, E-Store maintains an elastic number of partitions of arbitrary sizes, where each partition is assigned to a single core on some node in a computing cluster.

E-Monitor

E-Monitor periodically collects coarse-grained information about each node in the cluster, including CPU and memory utilization. When these metrics exceed a threshold, E-Monitor turns on tuple tracking for a short time window to collect the access frequency per tuple.



E-Planner



Table&uples:![k_1,k_2,…,k_T]!

1. E/Monitor&den6fies&ot&uples&&&&&& &&&&and&heir&weights&

& $(k_n,\%w_n)\%(k_m,\%w_m)\%\%$
&
&
&

2.&ar66on&old&uples&nto& &&&equally&weighted&anges&

& $r_1=[k_1 \ldots k_n)\%\%$
& $r_2=[k_{n+1} \ldots k_m)\%$
& $r_3=[k_{m+1} \ldots k_T]\%$
&

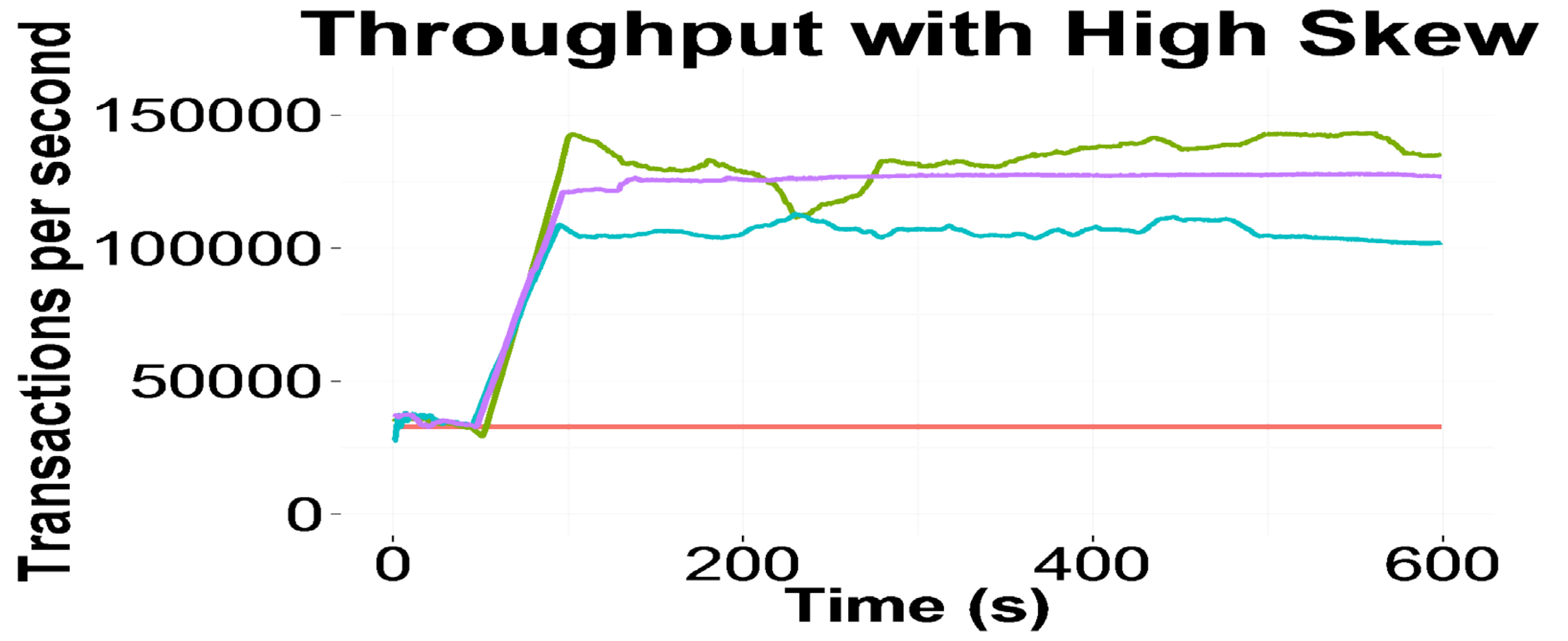
3.&Assign&ot&uples&o&ar66ons&

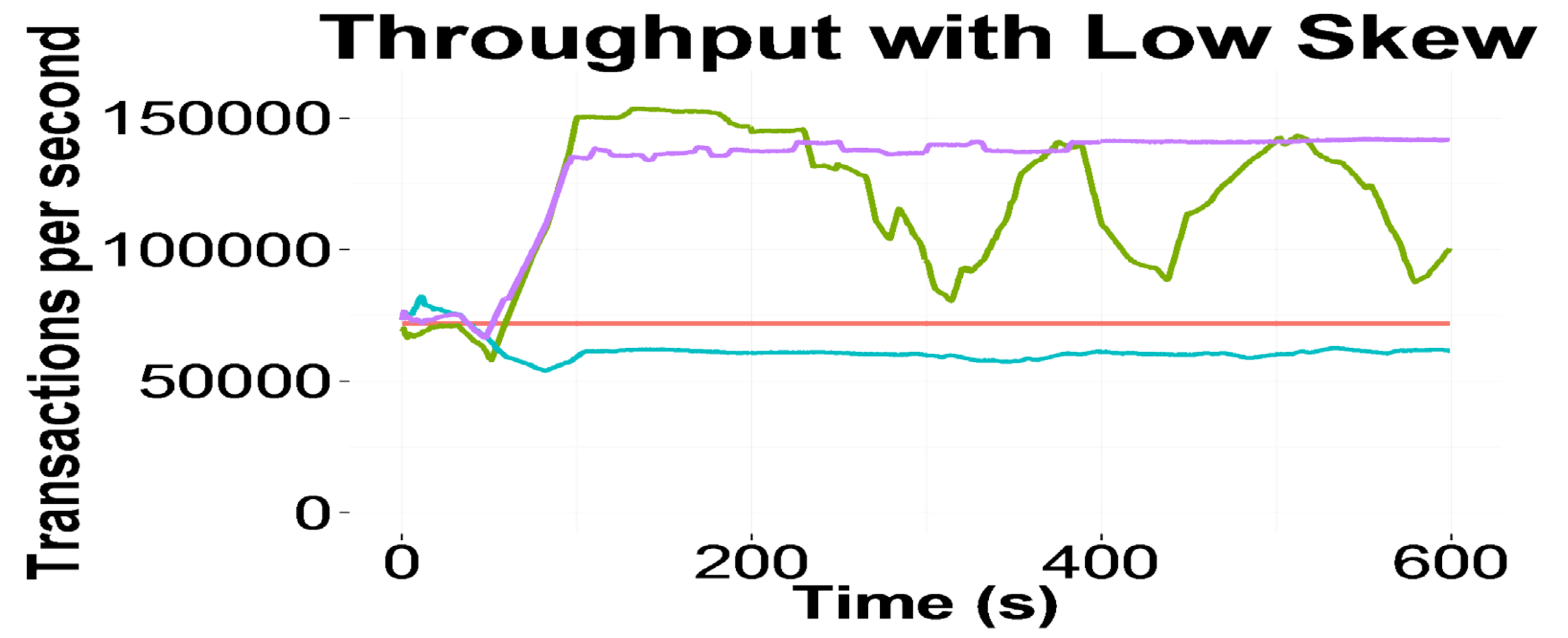
4.&Distribute&old&data&ver& &&&emaining&apacity&&&&
&
!

Experiments

We developed several different tuple placement planners. These planners aim to balance the workload between partitions. The planners perform differently based on load skew.

Planners
- Baseline
- First Fit
- Greedy
- Greedy Extended

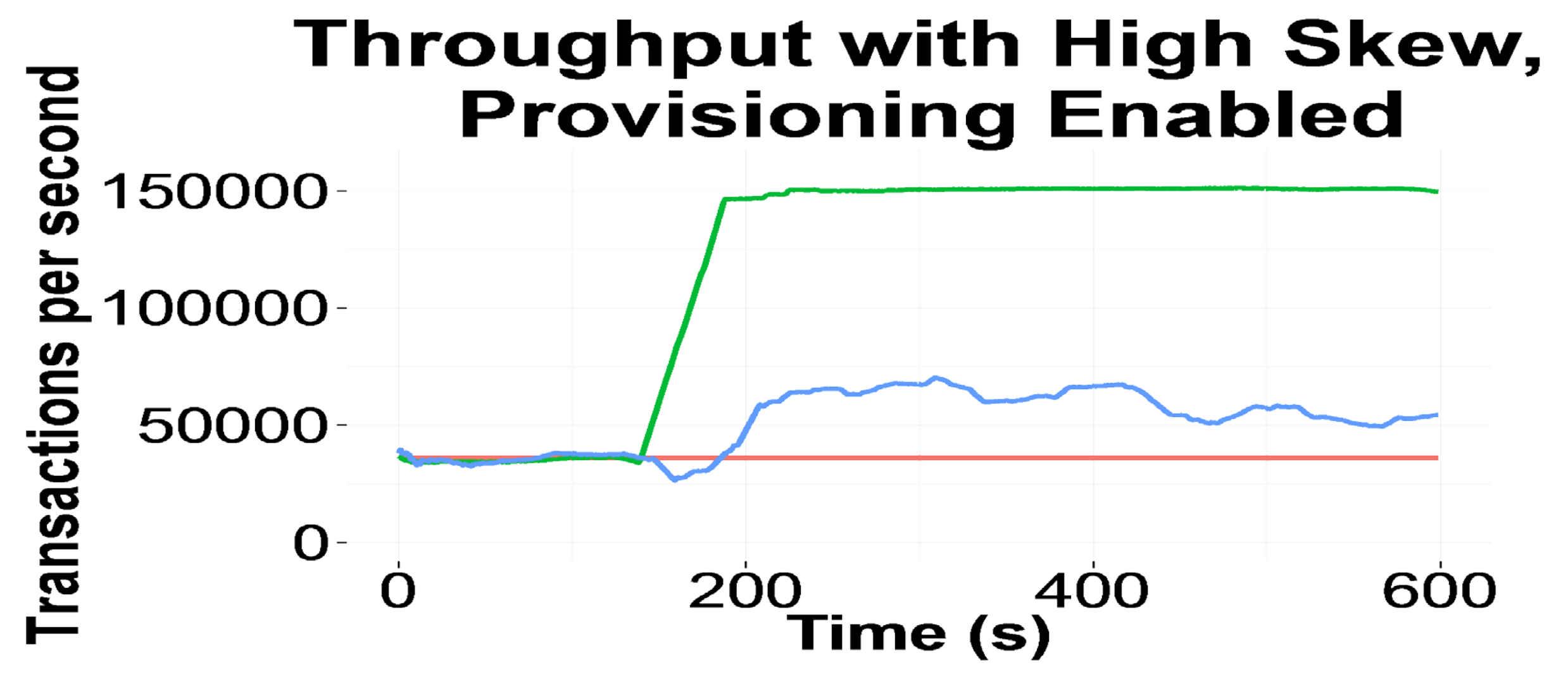




Greedy and Greedy Extended consider the locations of the tuples. Therefore, they migrate fewer tuples than First Fit, which suffers a drop in throughput due to migration cost. Greedy only moves hot tuples, so does not perform well with low skew. Greedy Extended moves both hot and cold tuples, and leads to significant performance improvements in terms of throughput and latency for all skew types.

Here we compare 1- and 2-tier placement with the Greedy Extended algorithm when scaling from 5 nodes to 10 nodes.

Planners
- Baseline
- Greedy Extended
- Greedy Extended One Tiered



2-tier Greedy Extended (GE) has the ability to split hot spots at the granularity of individual tuples. Therefore, 2-tier GE significantly improves the throughput for skewed workloads with minimal migration overhead compared to 1-tier GE.