Towards Balanced Allocations for DHTs

George Tsatsanifos (NTUA)
Vasilis Samoladas (TUC)

22nd Conference on Database and Expert Systems Applications
DEXA 2011, Toulouse, France
September 1, 2011

Outline

Peer-to-Peer Networks

Structured vs. Unstructured

Unstructured P2P

- Arbitrary links
- Routing through:
  - Flooding
  - Random walks
- Query resolution?

Structured P2P

- Employment of a globally consistent protocol
- Routing through network structure
- Guaranteed query resolution

Distributed Hash-Tables

In essence...

- Provision of a scalable lookup service.
- Consistent hashing variants assign (key,val) pairs to peers.

Contemporary DHTs support/approximate

1. Exact Queries (lookups)
2. Range Queries ($\leq d$)
3. Nearest Neighbors (K-NN)
4. Aggregation (gossiping min, max, avg, sum)

Balls-in-Bins Models

Principles

- In the standard model m balls are placed sequentially into n bins.
- Each ball probes d bins selected with equal probability and is placed in the least loaded bin.

Balls-in-Bins Models

Maximum Load

- Refers to the maximum number of balls in any bin at the end.
- Assumed that all balls weight the same...
- For $d = 1$, max-load is $(1 + o(1))\frac{logn}{loglogn}$.
- For $d > 1$, max-load is $\frac{loglogn}{logd} + O(1)$.
  as $n \to \infty$, for $m \geq n$, max-load is $(1 + o(1))\frac{loglogn}{logd} + \Theta\left(\frac{m}{n}\right)$.
- Holds even if bins are not selected indepedently
  as long as they are associated in $\Omega(logn)$ degree.

Problem Formulation

Load Balancing

- An optimization problem.
- Related to *NP-complete problems* and *combinatorics*.
- Approached through heuristics and approximate solutions.
- Practical extentions on many system parameters,
  e.g., optimal resource utilization, throughput, response time.
- Task-skew is the most severe impediment to network performance.

Problem Formulation

Objectives

- Our intention is to develop a realistic load-balancing scheme, which will be used on top of any overlay.
- We do not want to damage any overlay properties, e.g., efficient complex query support.

Course of Action

We consider protocols based on migration in order to improve scalability.

The Virtual Nodes Technique

Characteristics

- Associating keys with virtual nodes
- Mapping multiple virtual nodes with unrelated identifiers to each peer.
- Intuitively, provides a more uniform coverage of the identifier space.
- Average latency increases due to longer paths in the larger overlay.
- Originally proposed to handle data-skew imbalances!

The Virtual Nodes Technique

Applying the Balls-and-Bins Model

- Assume there are n peers and $m = \alpha n$ nodes, where $\alpha \geq 1$.
- Consider peers as bins and nodes as balls.
- We associate each node with a peer-set consisted of its host and its host's linked hosts.
- There are overlapping bin-sets.

The "Enhanced" Virtual Nodes Technique...

Our Augmentation Focusing on Task-Skew

The Inflationary Balancing Algorithm

1. When a peer p exceeds a load threshold T or a time-window is elapsed
2. A balancing process is triggered among all peers in p's neighborhood.
3. Peers with load above the average make available their heaviest node(s) whose loads sum to that specific load gap.
4. Will be assigned to the underloaded peers through a greedy process.
5. Each dispatched node repeatedly migrates to the lightest host, starting from the heaviest node.
6. We repeat this process for another heavy host.

The Replication Technique

Principles

- A popular key-space area becomes available through many peers.
- When copying hotspots, only portion of the node traffic reaches each replica, alleviating heavy peers (bottlenecks)
- ...at the cost of a storage overhead.

The Replication Technique

Applying the Balls-and-Bins Model

- Assume there are n peers and $m = \beta n$ nodes, $0 < \beta \leq 1$.
- Consider nodes as bins and peers as balls.
- Each node is made available through many peers.
- We associate each peer with a node-set consisted of the node it replicates and its links.
- There are overlapping bin-sets.

The "Enhanced" Replication Technique

Our Modification

The Deflationary Balancing Algorithm

1. We choose a peer p due to its heavy load.
2. A balancing procedure is initiated among all peers of p's neighborhood.
3. Assign repeatedly to each peer a replica of the heaviest replica of the neighborhood.
4. We repeat this process for another heavy replica.
5. Greedy approach

The Multiple Realities Technique

Principles

- Maintaining multiple, independent coordinate spaces with each peer being assigned a different zone in each one.
- Contents of the hash table are replicated on every reality, and thus, improving data availability and fault-tolerance.
- Originally targets at ameliorating latency.

The Multiple Realities Technique

How it was designed to work...

- A peer "broadcasts" his request through his nodes in all realities.
- The fastest answer is returned to the user.

However, latency is very slightly affected, especially for range queries!

But we changed that for our purposes...

- A peer enacts each of his queries in a randomly selected reality.
- We take advantage of the enhanced data availability.

The Multiple Realities Technique

Our Augmentation Focusing on Task-Skew

The Pseudo-Inflationary Balancing Algorithm

1. We choose a peer p due to its heavy load.
2. A balancing procedure is initiated among the peers in p's neighborhood.
3. Assign sequentially the heaviest node in reality j to the lightest peer, where peer-loads are computed for the $1, \cdots, j-1$ previous realities.
4. Ties are broken arbitrarilly.

The Multi-Bin Adaptive Balancing Protocol

Motivation

So far, we have managed to enhance existing techniques.
But we want more...

Keystone

- Combine the previous schemes into a single balancing protocol.
- Parts of the network might use the virtual nodes technique while... some nodes might be replicated by many hosts at the same time.
- Thus, exploit the benefits from each technique!
- Add dynamicity... a lot of!

So, after the previous prerequisite study we are in position of realizing this.

The Multi-Bin Adaptive Balancing Protocol

Protocol Specifications

- Assume r overlays, each consisted of n_r virtual nodes.
- Both parameters are fixed.
- New node replicas can be created when new peers join.
- Node replicas may be deleted when a peer departs.
- Data cannot be lost.

The Multi-Bin Join Procedure

Peer p wants to join the network.

1. Peer p contacts an already pariticipating peer q of the network.
2. Peer q assigns to p, starting from the heaviest, nodes from all his known peers and himself until its load equals to $\frac{total\ load}{peers+1}$.
   - Nodes that migrate change ownership for the new peer only.
   - A new replica is created and assigned to peer p when the owner contains a sole node. Then, both peers contain a copy of that node.
3. This procedure is repeated for every reality.

The Multi-Bin Departure Procedure

Peer p wants to depart from the network.

1. Peer p is the only peer making available nodes.
2. Peer p affiliates the lightest peer he knows with his unique replicas, starting from the heaviest.
3. Again, this procedure takes place for all realities.

Experimental Analysis

Setting

Static Simulations

- Consist of $10K$ bins and $g \times 10K$ balls, where $g \in \{1, 2, ..., 12\}$.
- Compare three different policies:
  - **Naive** represents a straightforward random assignment of balls to bins.
  - **Local** represents our prudent mechanisms for each ballancing technique.
  - **Ideal** for a centralized balancing mechanism.
- We make use of the PGrid-Z overlay.
- Used *real spatial* and *synthetic* datasets.
- Datasets consist of $1M$ keys.
- Querysets consist of $50K$ range queries.
- Each query evaluates 50 tuples.

Experimental Analysis

Setting

Dynamic Simulations

- Our experiments simulate a dynamic environment.
- They consist of three stages, *growing*, *steady*, and *shrinking* stage.
- *Real spatial* and *synthetic* high-dimensional datasets for comparison.
- We initiate a network of $1.25K$ hosts, increasing to $80K$.

Evaluation Metrics

Load-Centric Performance Metrics

Maximum Throughput

Modelling a peer by an GI/GI/1/P queue

- maximum service rate γ_i is given for each peer $1 \leq i \leq n$
- a set P of processes with associated popularities $\phi_p \forall p \in P$
- m_i^p denotes the number of messages received by i in process p.
- $E[s]$ denotes the expected service demand of an incoming message.

$$\Lambda_{\max} = \frac{1}{E[s] \ \max_i \sum_{p \in P} \phi_p \, m_i^p}$$

Evaluation Metrics

Task-Load Fairness Metrics

Fairness Index

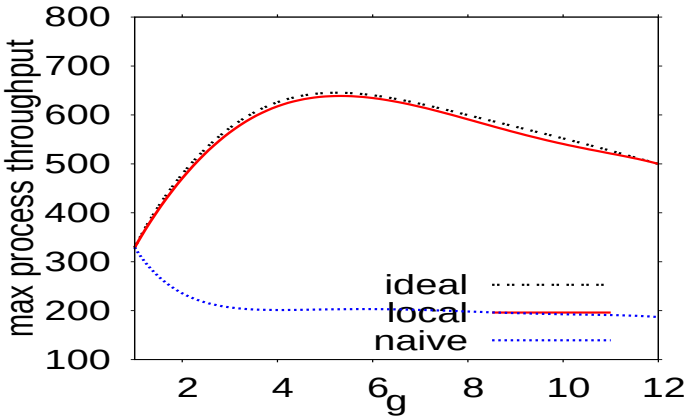
- Population Size Independence
- Scale and Metric Independence
- Boundedness
- Continuity

$$f(x) = \frac{(\sum_{i=1}^{N} x_i)^2}{n \sum_{i=1}^{N} x_i^2}, \text{ where } x_i \geq 0$$

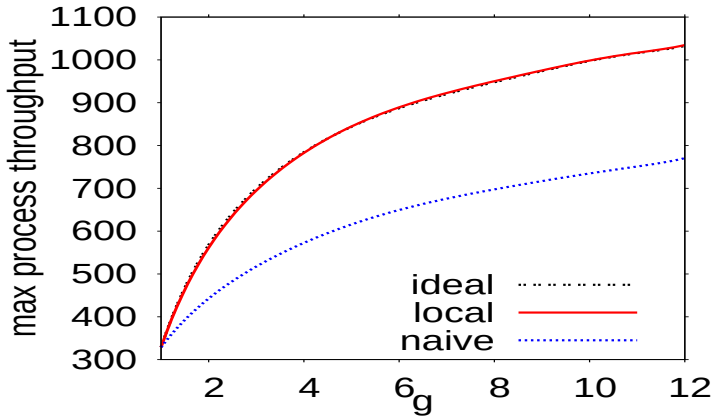
Results

Performance Evaluation

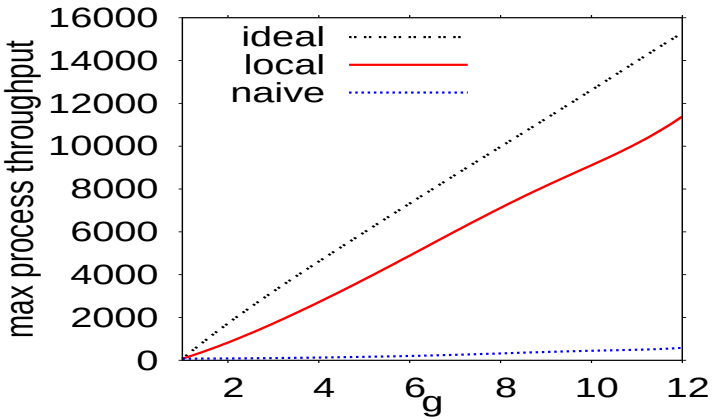
(a) virtual nodes



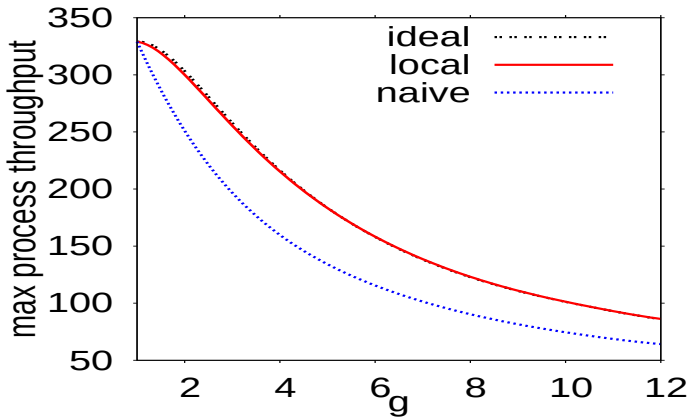
(b) independent realities



(c) replication



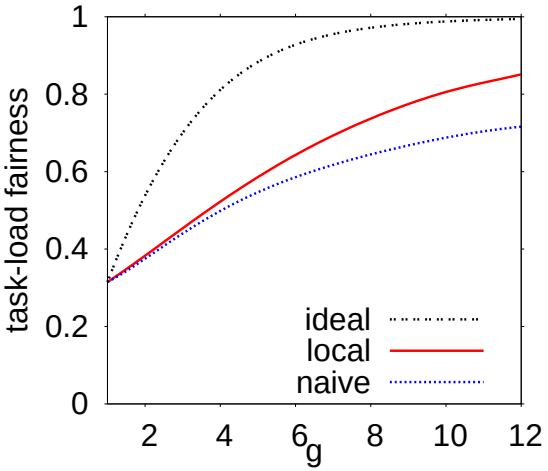
(d) multiple realities



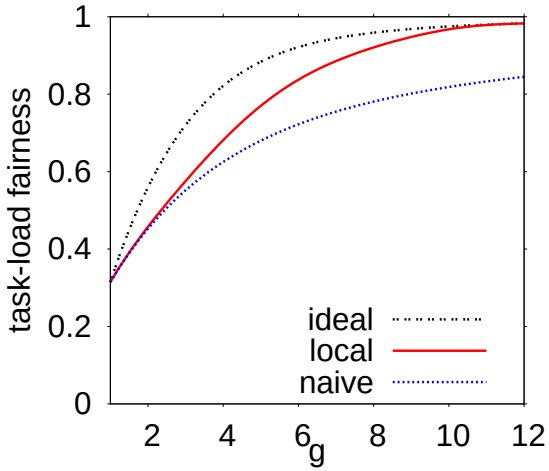
Results

Task-Load Fairness

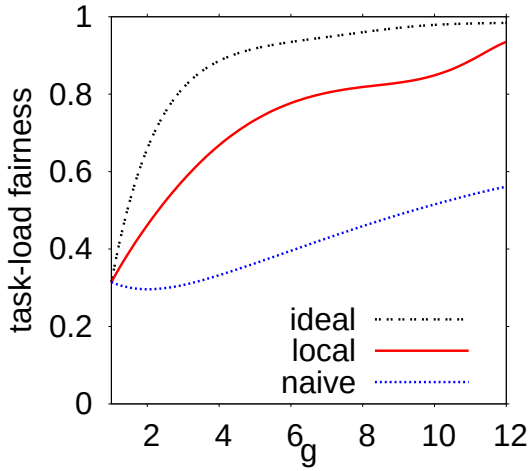
(a) virtual nodes



(b) independent realities



(c) replication

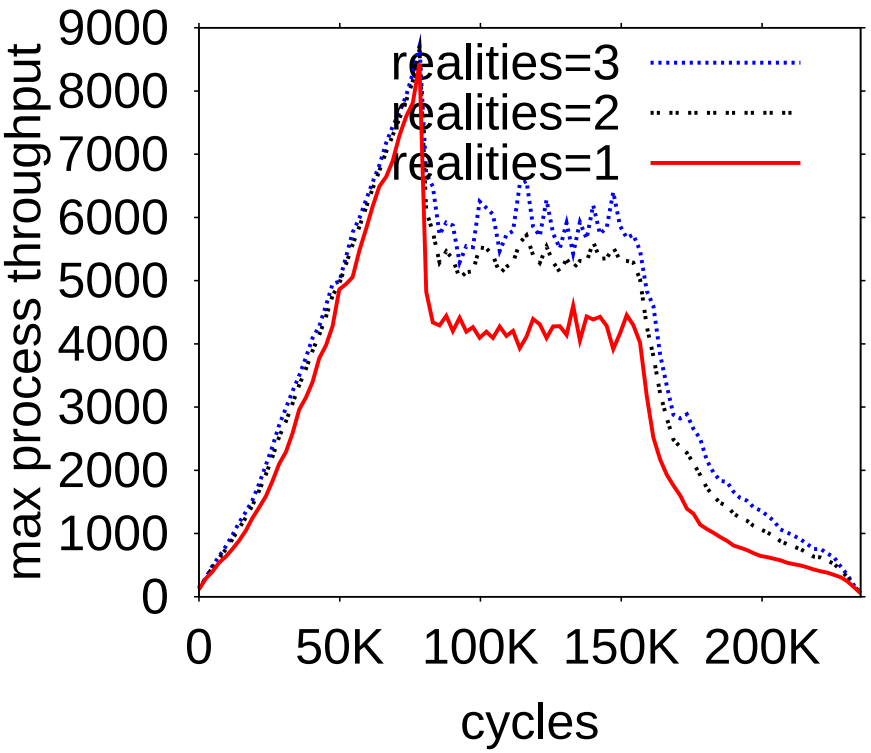


Results

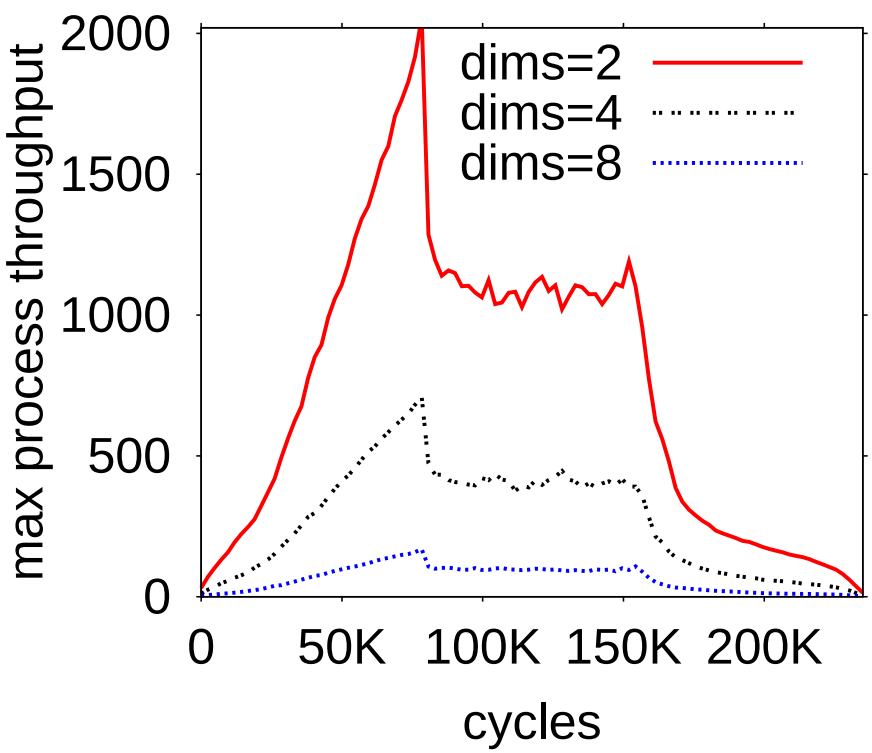
The Multi-Bin Adaptive Balancing Protocol

Performance Evaluation

(a) spatial



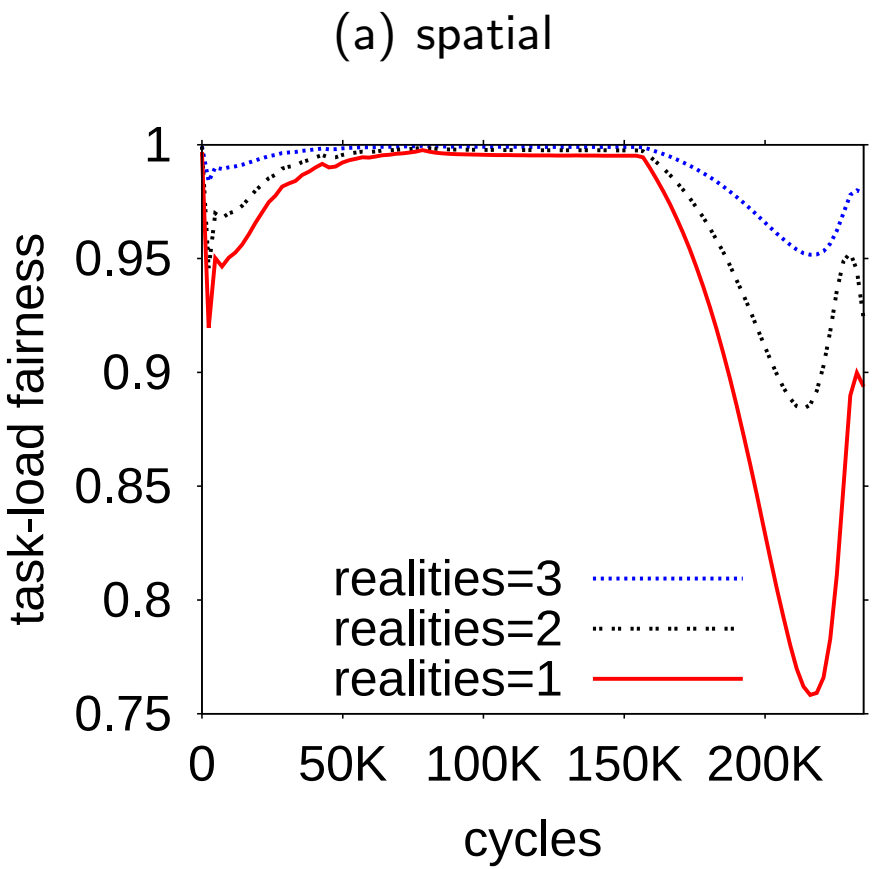
(b) multi-dimensional

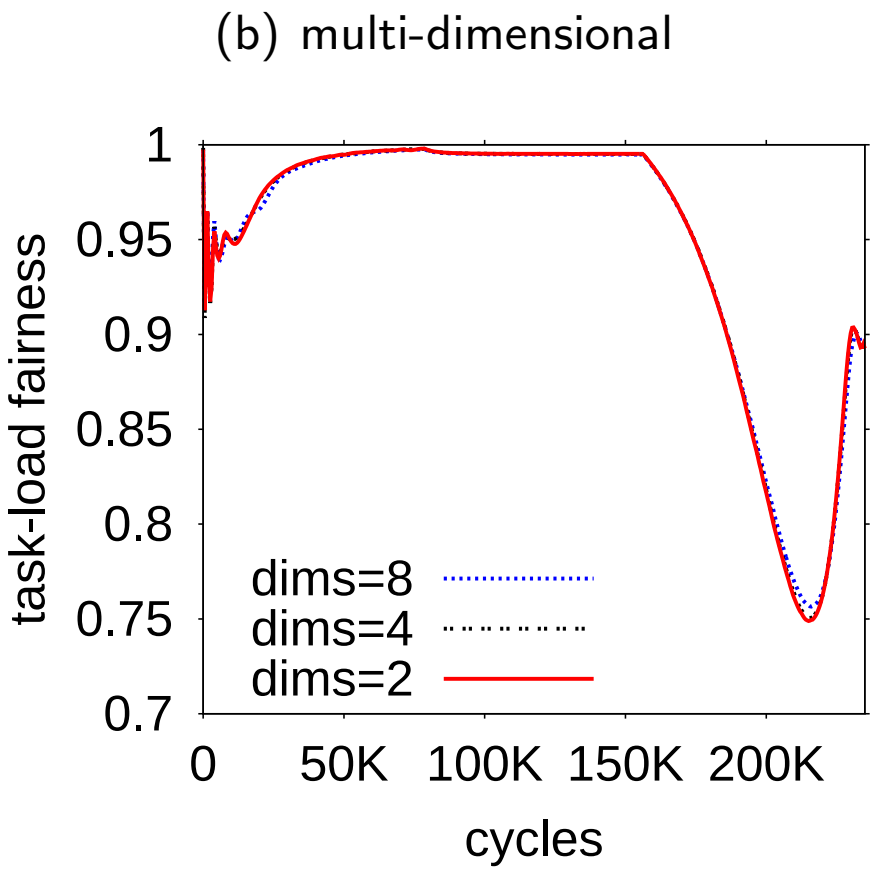


Results

The Multi-Bin Adaptive Balancing Protocol

Task Load Fairness



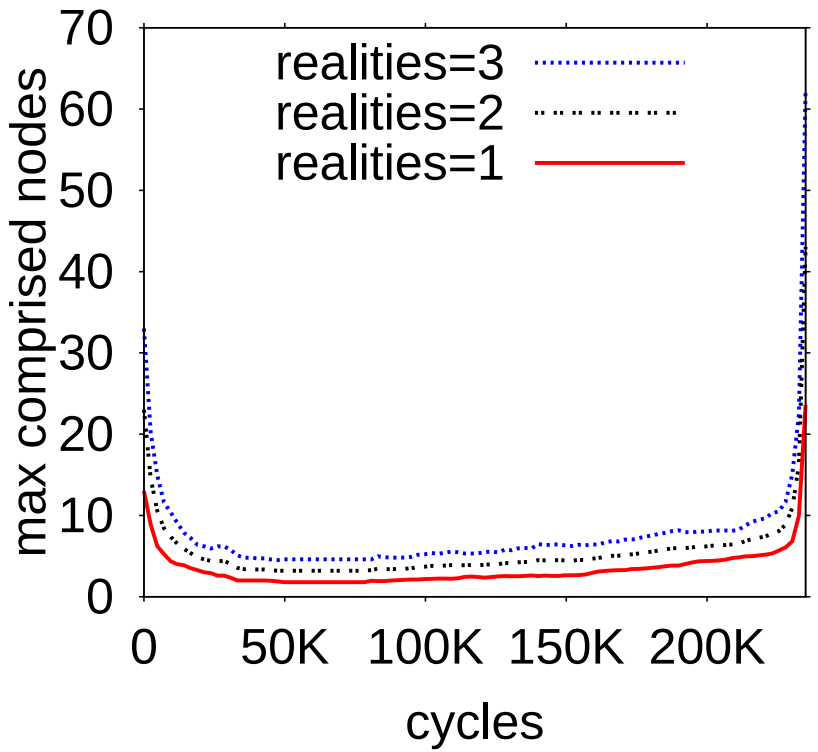


Results

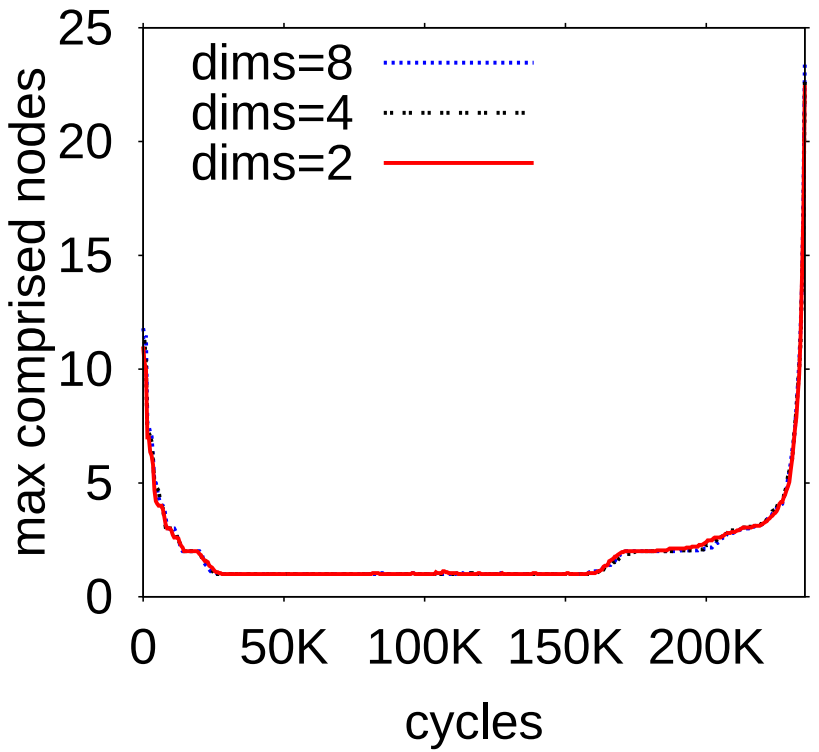
The Multi-Bin Adaptive Balancing Protocol

Maintenance Cost

(a) spatial



(b) multi-dimensional

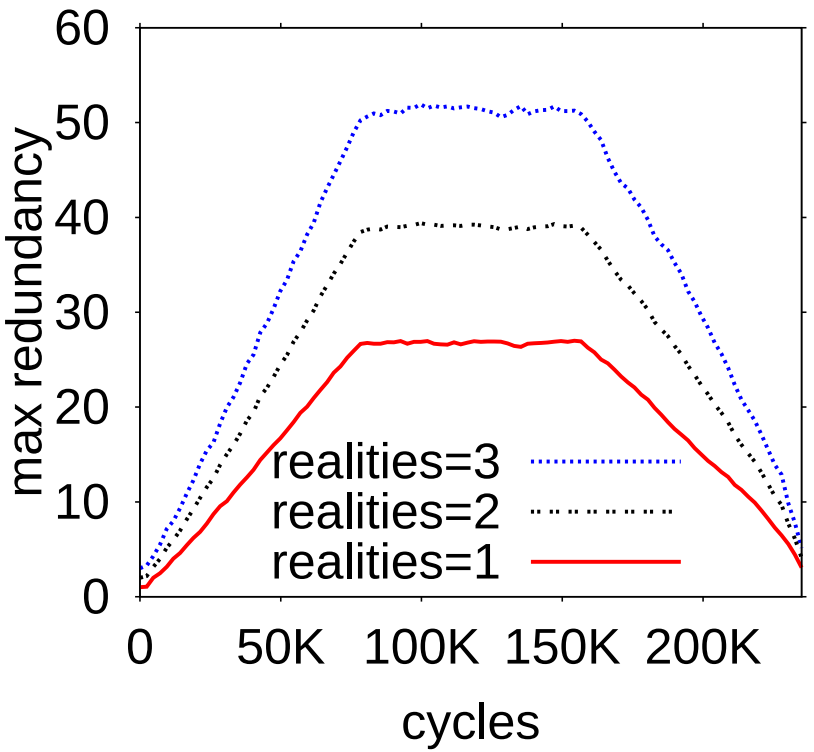


Results

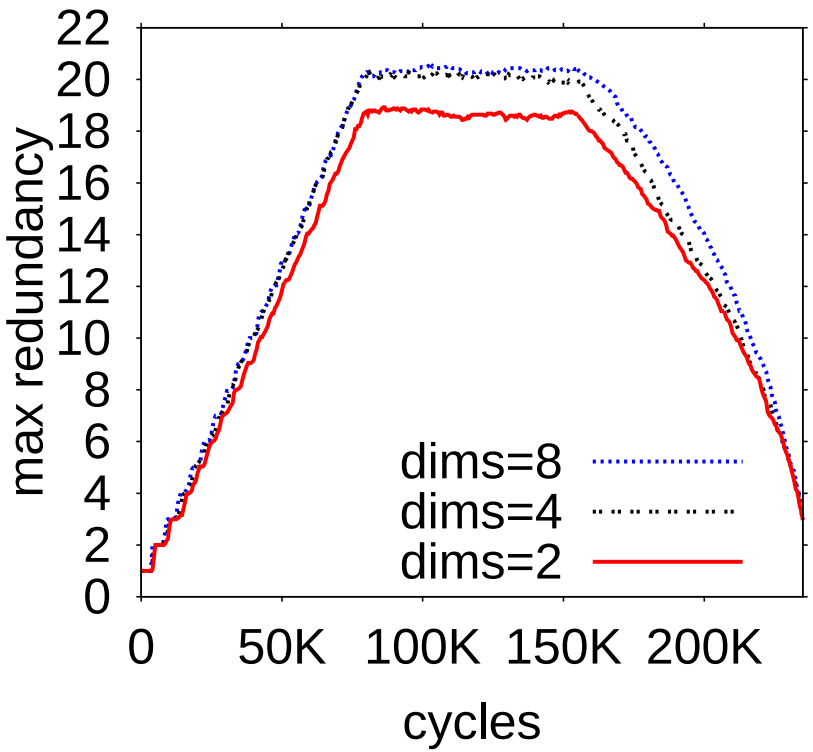
The Multi-Bin Adaptive Balancing Protocol

Redundancy

(a) spatial



(b) multi-dimensional

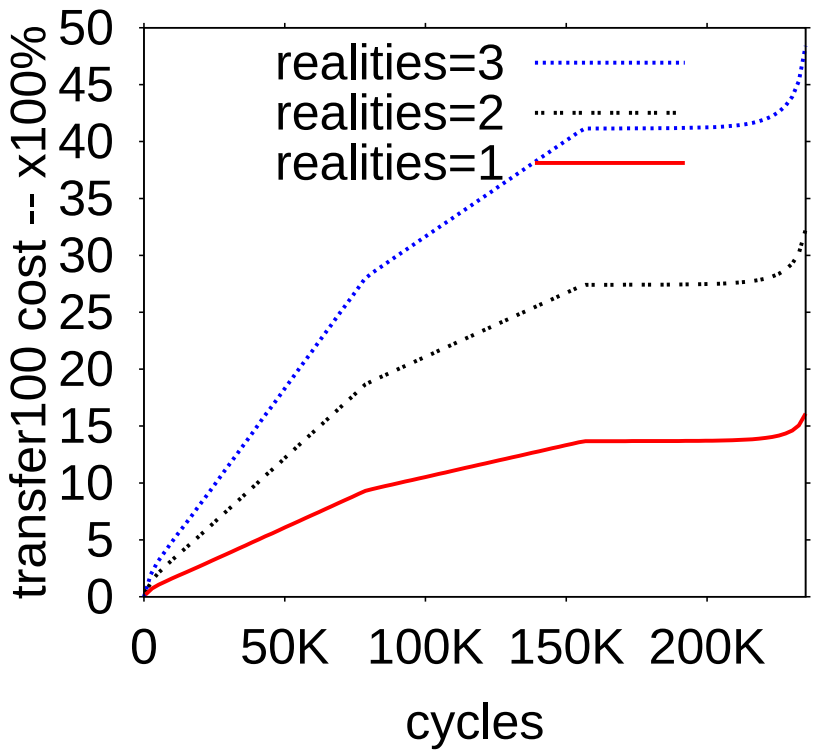


Results

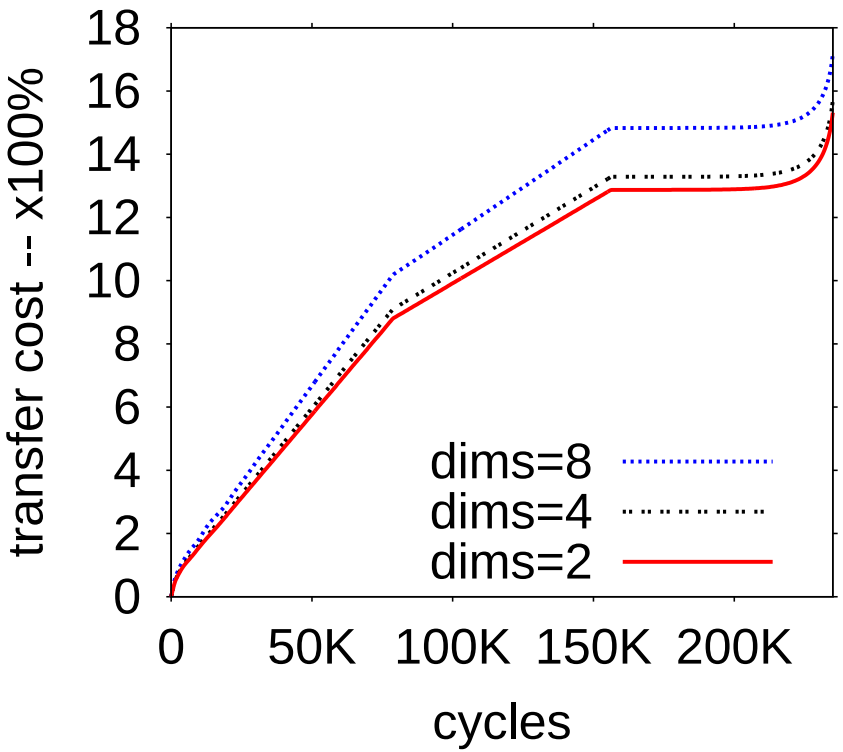
The Multi-Bin Adaptive Balancing Protocol

Bandwidth

(a) spatial



(b) multi-dimensional



The Aftermath

Virtual Nodes

- Enhanced performance devoid of additional redundancy.
- Accumulates data load from all containing nodes.
- Preserving multiple structures in each peer.
- Multiple maintenance cost for each peer.

Replication

- Extremely efficient and flexible.
- Capable of adapting to any skewness
- Costly when some nodes become over-replicated.

The Aftermath

Parallel Realities

- Appropriate exclusively for lookups.
- Fixed redundancy equal to the number of realities.
- Accumulating data load from all realities.

Independent Realities

- Useful even for naive allocations.
- Lacking flexibility.
- Suits for high-dimensionality cases or non-steady skewness.

The Aftermath

The Multi-Bin Balancing Protocol

Simple based on well-known heuristics of the makespan problem.

Flexible empowers the cooperation of different balancing principles.

Adaptive blunts imbalances of arbitrary load distribution.

Dynamic performed on-line without disrupting overlay operations.

Efficient converges quickly.

Realistic and easy to implement for all of the above!

Our paradigm

- can be incorporated on top of existing overlays.
- works even for hetergoneous peer networks, as long as a task-load function is provided.

Questions?