Law, Science and Technology
MSCA ITN EJD n. 814177

RIoE
Rights of Internet of Everything

Towards Decentralized Complex Queries over Distributed Ledgers: a Data Marketplace Use-case

Mirko Zichichi[1,2], Luca Serena[2], Stefano Ferretti[3], Gabriele D'Angelo[2]

[1]Universidad Politécnica de Madrid
[2]University of Bologna
[3]University of Urbino "Carlo Bo"

Overview

Introduction

Distributed Ledger Technologies (DLT) and Decentralized File Storages (DFS)

DLT and DFS are being increasingly used to create **common, decentralized and trustless infrastructures** where participants interact and collaborate in Peer-to-Peer interactions. They enable:

Distributed Ledger Technologies (DLT) and Decentralized File Storages (DFS)

DLT and DFS are being increasingly used to create **common, decentralized and trustless infrastructures** where participants interact and collaborate in Peer-to-Peer interactions. They enable:

- secure transactions between **untrusted parties** through consensus mechanisms

Distributed Ledger Technologies (DLT) and Decentralized File Storages (DFS)

DLT and DFS are being increasingly used to create **common, decentralized and trustless infrastructures** where participants interact and collaborate in Peer-to-Peer interactions. They enable:

- secure transactions between **untrusted parties** through consensus mechanisms
- high data **availability**

Distributed Ledger Technologies (DLT) and Decentralized File Storages (DFS)

DLT and DFS are being increasingly used to create **common, decentralized and trustless infrastructures** where participants interact and collaborate in Peer-to-Peer interactions. They enable:

- secure transactions between **untrusted parties** through consensus mechanisms
- high data **availability**
- ability to automate and enforce processes (through smart contracts)

Query Issues

1) data stored in DLTs and DFS are usually **unstructured** and need to be **filtered and indexed** before any **complex query**

Query Issues

1) data stored in DLTs and DFS are usually **unstructured** and need to be **filtered and indexed** before any **complex query**

2) there are **no diffused efficient mechanisms to query** a certain type of data, that do not involve **centralization** (e.g. index data in a central database)

Our work

- System for the search of data in DLTs and DFS according to their content or meaning

Our work

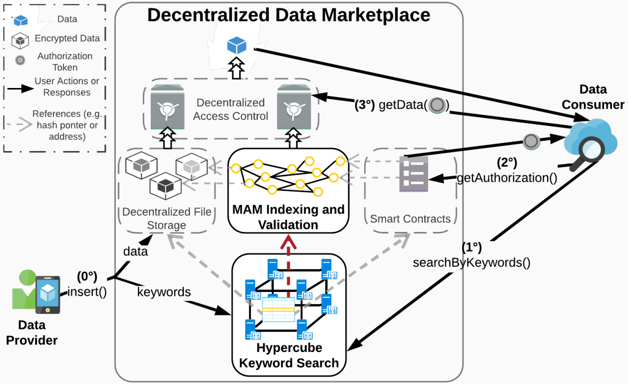
- System for the search of data in DLTs and DFS according to their content or meaning
- **Distributed Hash Table (DHT)** as a layer placed over DLTs: DHT → distributed data structure that maps “**keys**” into “**values**”.

Our work

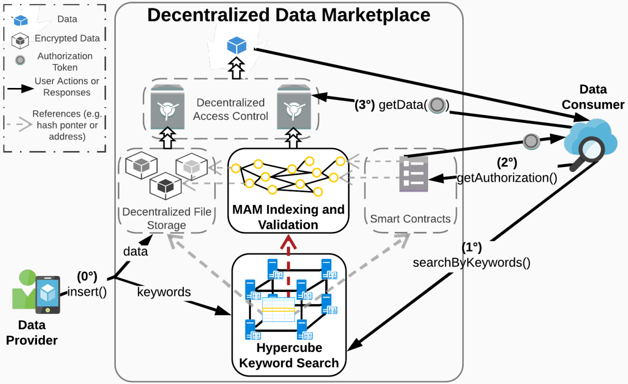
- System for the search of data in DLTs and DFS according to their content or meaning
- **Distributed Hash Table (DHT)** as a layer placed over DLTs: DHT $\rightarrow$ distributed data structure that maps "**keys**" into "**values**".
- **Hypercube** to organise the topological structure of such a DHT network.

Use Case

Decentralized Data Marketplace Use Case

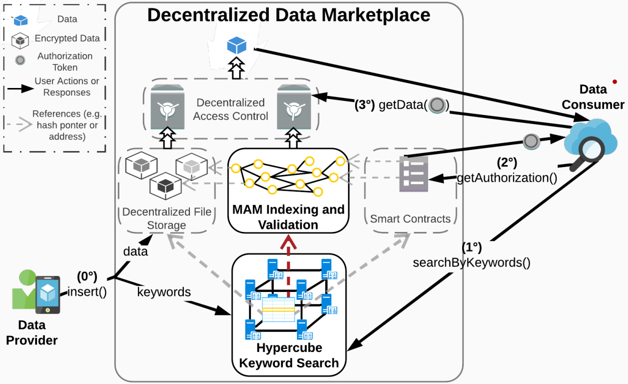


Decentralized Data Marketplace Use Case



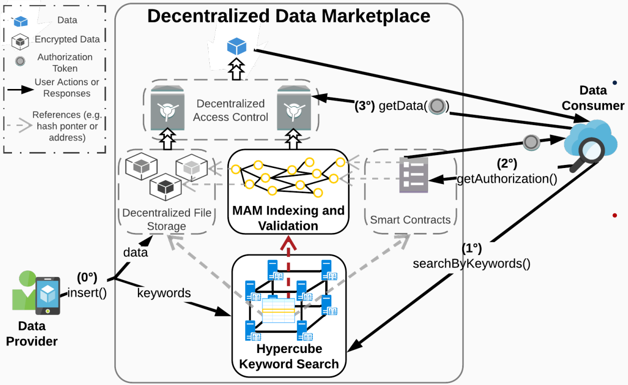
- **DFS** → used to store data in an encrypted form, offering high availability

Decentralized Data Marketplace Use Case



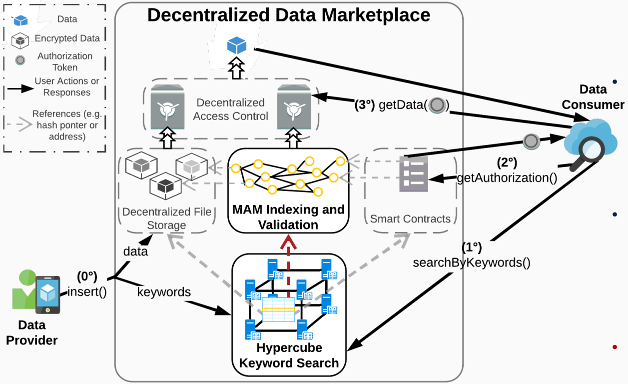
- **DFS** → used to store data in an encrypted form, offering high availability
- A **decentralized access control system** → to get the data from the DFS once they have been authorized

Decentralized Data Marketplace Use Case



- **DFS** → used to store data in an encrypted form, offering high availability
- A **decentralized access control system** → to get the data from the DFS once they have been authorized
- **Smart contracts** have the ability to provide a distributed authorization mechanism following a policy

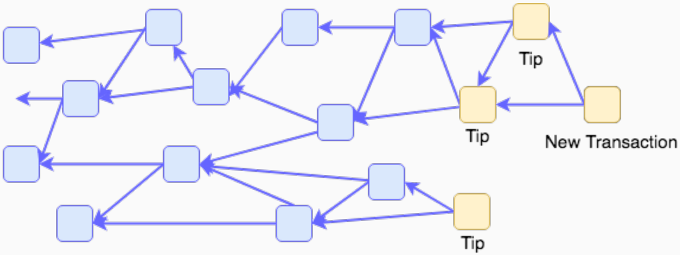
Decentralized Data Marketplace Use Case



- **DFS** → used to store data in an encrypted form, offering high availability
- A **decentralized access control system** → to get the data from the DFS once they have been authorized
- **Smart contracts** have the ability to provide a distributed authorization mechanism following a policy
- A **DLT** such as **IOTA** enable the data indexing and validation (in form of hash pointers)

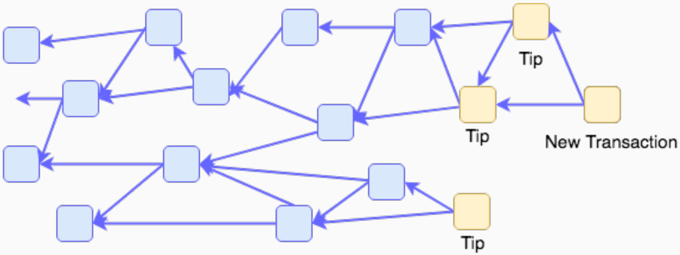
IOTA Masked Authentication Messaging Channels

- **IOTA** → network of nodes that holds a distributed ledger where transactions are validated without fees



IOTA Masked Authentication Messaging Channels

- IOTA → network of nodes that holds a distributed ledger where transactions are validated without fees
- Masked Authenticated Messaging (MAM) → communication protocol that adds the functionality to emit and access an encrypted data channels over IOTA



MAM Channels and Data Retrieval

- To obtain information from a message within a MAM channel, it is necessary to know the exact address of the message or of the channel, i.e. the **root value**

MAM Channels and Data Retrieval

- To obtain information from a message within a MAM channel, it is necessary to know the exact address of the message or of the channel, i.e. the **root value**
- *QEZXKW9HOPYNUGPNLOBXKZJEI9UJTNTACFVFNLYLX*

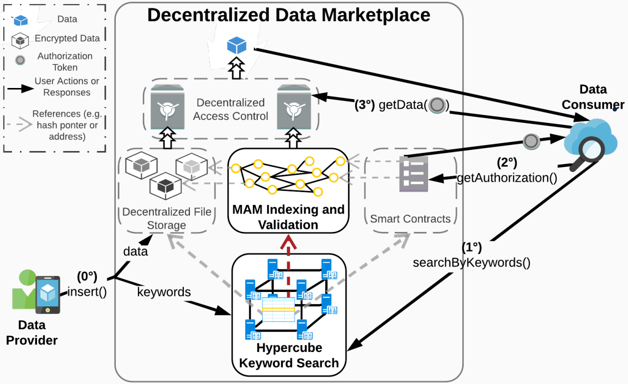
MAM Channels and Data Retrieval

- To obtain information from a message within a MAM channel, it is necessary to know the exact address of the message or of the channel, i.e. the **root value**
- *QEZXKW9HOPYNUGPNLOBXKZJEI9UJTNTACFVFNLYLX*
- this root, and in general DLT addresses, **do not provide any information** related to the type and kind of data

MAM Channels and Data Retrieval

- To obtain information from a message within a MAM channel, it is necessary to know the exact address of the message or of the channel, i.e. the **root value**
- *QEZXKW9HOPYNUGPNLOBXKZJEI9UJTNTACFVFNLYLX*
- this root, and in general DLT addresses, **do not provide any information** related to the type and kind of data
- in our system every single message is indexed by a **keyword set**, that is then exploited to search for specific kinds of contents ⇒

Decentralized Data Marketplace Use Case

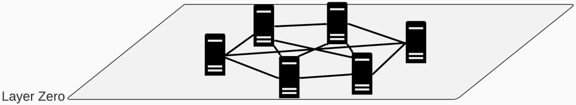


A distributed mechanism for the search of data is in charge of associating keywords to addresses or references stored in DLTs, smart contracts and DFS.

Hypercube DHT

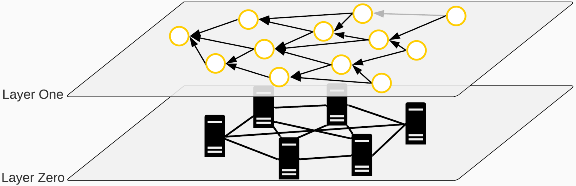
Layer Two Lookup Scheme

- DLT **P2P Network**



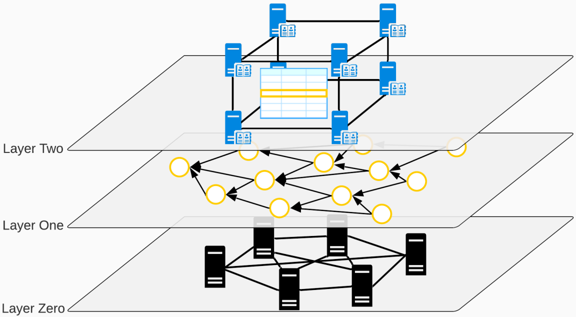
Layer Two Lookup Scheme

- DLT **P2P Network**
- Data are stored in a DFS and/or referenced in a **IOTA MAM Channels**.



Layer Two Lookup Scheme

- DLT **P2P Network**
- Data are stored in a DFS and/or referenced in a **IOTA MAM Channels**.
- Layer two solution → MAM messages associated to a **keyword set** in a DHT



Keywords Sets

- $O \leftarrow$ set of all MAM messages in IOTA

Keywords Sets

- **O** $\leftarrow$ set of all MAM messages in IOTA
- DHT for mapping **o** $\in O$ **to a keyword set K**$_o \subseteq W$ (W is the keyword space)

Keywords Sets

- **O** $\leftarrow$ set of all MAM messages in IOTA
- DHT for mapping **o** $\in O$ **to a keyword set K**$_o \subseteq W$ (W is the keyword space)
- By using a **uniform hash function**
  $h : W \rightarrow \{0, 1, \ldots, r - 1\}$
  K can be represented by a **string of bits u** $\rightarrow$ 101001

Keywords Sets

- **O** $\leftarrow$ set of all MAM messages in IOTA
- DHT for mapping **o** $\in O$ **to a keyword set** $\mathbf{K}_o \subseteq W$ (W is the keyword space)
- By using a **uniform hash function**
  $h : W \rightarrow \{0, 1, \ldots, r - 1\}$
  K can be represented by a **string of bits u** $\rightarrow$ 101001
- in **u the 1s are set in the positions** given by
  $one(u) = \{h(k) \mid k \in K\}$

Keywords Sets

- **O** $\leftarrow$ set of all MAM messages in IOTA
- DHT for mapping **o** $\in O$ **to a keyword set** $\mathbf{K}_o \subseteq W$ (W is the keyword space)
- By using a **uniform hash function**
  $h : W \rightarrow \{0, 1, \ldots, r-1\}$
  K can be represented by a **string of bits u** $\rightarrow$ 101001
- in **u the 1s are set in the positions** given by
  $one(u) = \{h(k) \mid k \in K\}$
- E.g.: $o = MAM\ msg\ indexed\ by\ QEZ...OBX\ root,\ K = \{temperature, celsius\}$
  $h(temperature) = 3, h(celsius) = 5$
  K is represented by $u = 000101 \Rightarrow$ **DHT stores (000101,*QEZ...OBX*)**

Hypercube based DHT

- We use these r-bit strings to identify logical nodes in a **DHT network**

Hypercube based DHT

- We use these r-bit strings to identify logical nodes in a **DHT network**
- network topology $\rightarrow H_r(V, E)$ r-dimensional **hypercube**

Hypercube based DHT

- We use these r-bit strings to identify logical nodes in a **DHT network**
- network topology $\rightarrow H_r(V, E)$ r-dimensional **hypercube**
- V set of vertices that represent **logical nodes**

Hypercube based DHT

- We use these r-bit strings to identify logical nodes in a **DHT network**
- network topology $\rightarrow H_r(V, E)$ r-dimensional **hypercube**
- **V** set of vertices that represent **logical nodes**
- **E** set of edges formed when two vertices differ of only one bit (they are also network **neighbors**), e.g. 1011 and 1010.

Hypercube based DHT

- We use these r-bit strings to identify logical nodes in a **DHT network**
- network topology $\rightarrow H_r(V, E)$ r-dimensional **hypercube**
- **V** set of vertices that represent **logical nodes**
- **E** set of edges formed when two vertices differ of only one bit (they are also network **neighbors**), e.g. 1011 and 1010.
- to find out how far apart two vertices u and v are $\rightarrow \boldsymbol{HammingDistance}(u, v) = \sum_{i=0}^{r-1}(u_i \oplus v_i)$, $\oplus$ is the XOR operation and u_i is the bit at the i-th position.

Multiple Keywords Search

- **Pin Search** - $\{o \in O \mid K_o = K\}$
  gets all and only the objects associated with a keyword set K
  e.g. *pinSearch({temperature, celsius})* = (*000**101**,QEZ...OBX*), (*000**101**,IHU...9HZ*), ...

Multiple Keywords Search

- **Pin Search** - $\{o \in O \mid K_o = K\}$
  gets all and only the objects associated with a keyword set K
  e.g. *pinSearch({temperature, celsius})* = (*000**101**,QEZ...OBX*), (*000**101**,IHU...9HZ*), ...
- **Superset Search** - $\{o \in O \mid K_o \supseteq K\}$
  also gets objects that can be described by keywords sets that include K
  e.g. *superSetSearch({temperature, celsius})* = (*000**101**,QEZ...OBX*), (*000**111**,XTL...A9Z*), ...

Performance Evaluation

Test Setup

- **Simulated DHT network** (using PeerSim)

Test Setup

- **Simulated DHT network** (using PeerSim)
- **Nodes number** $\rightarrow$ from 128 ($r = 7$) up to 8192 ($r = 13$)

Test Setup

- **Simulated DHT network** (using PeerSim)
- **Nodes number** $\rightarrow$ from 128 ($r = 7$) up to 8192 ($r = 13$)
- Randomly created keywords-objects (i.e. MAM message roots) $\rightarrow$ **objects number 100, 1000 and 10000**

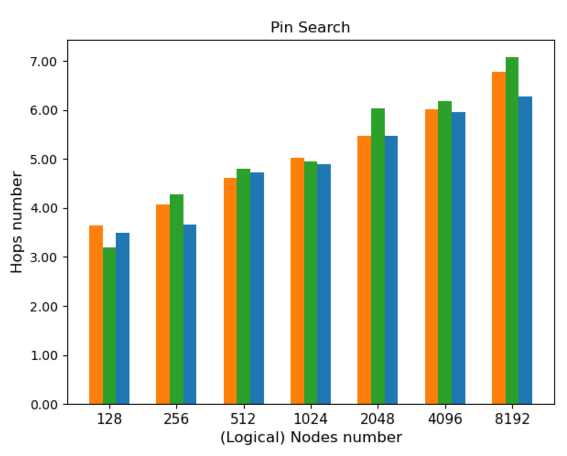
Test Setup

- **Simulated DHT network** (using PeerSim)
- **Nodes number** $\rightarrow$ from 128 ($r = 7$) up to 8192 ($r = 13$)
- Randomly created keywords-objects (i.e. MAM message roots) $\rightarrow$ **objects number 100, 1000 and 10000**
- We evaluated the **number of hops** required for each new query

Test Setup

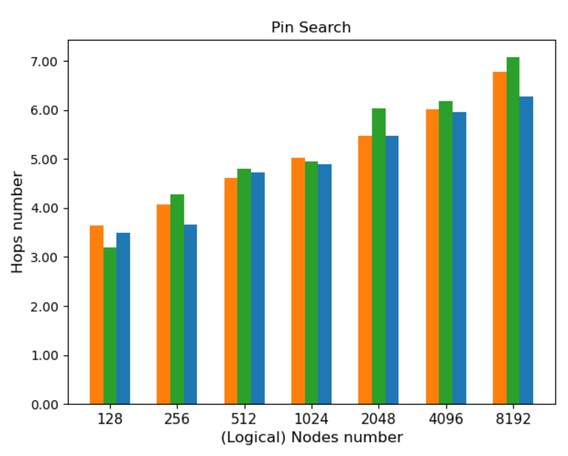
- **Simulated DHT network** (using PeerSim)
- **Nodes number** $\rightarrow$ from 128 ($r = 7$) up to 8192 ($r = 13$)
- Randomly created keywords-objects (i.e. MAM message roots) $\rightarrow$ **objects number 100, 1000 and 10000**
- We evaluated the **number of hops** required for each new query
- For each type of test $\rightarrow$ 50 repetitions

Pin Search Results



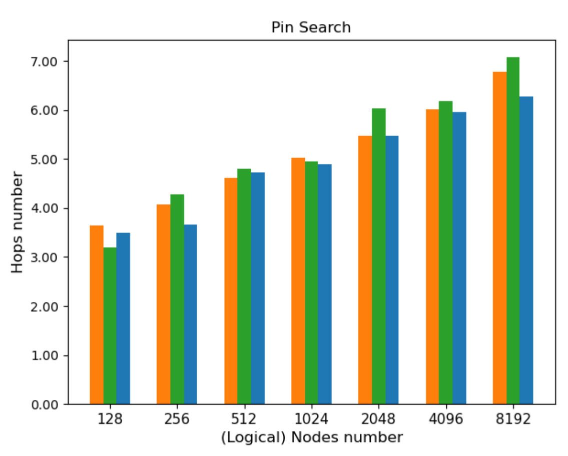
- Average number of hops increases

Pin Search Results



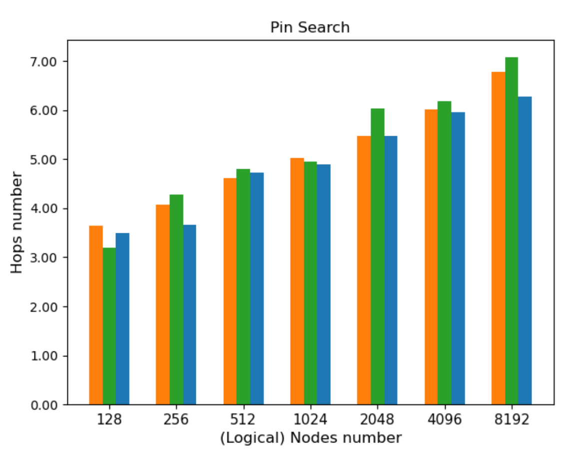
- Average number of hops increases
- from about 3.5 for 128 nodes ($r = 7$)

Pin Search Results



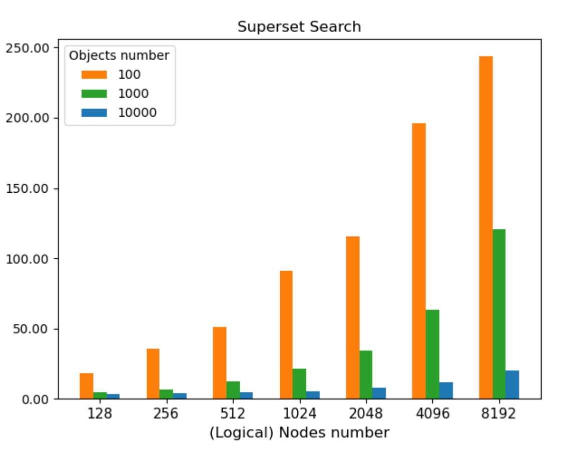
- Average number of hops increases
- from about 3.5 for 128 nodes ($r = 7$)
- to about 6.72 for 8192 nodes ($r = 13$)

Pin Search Results



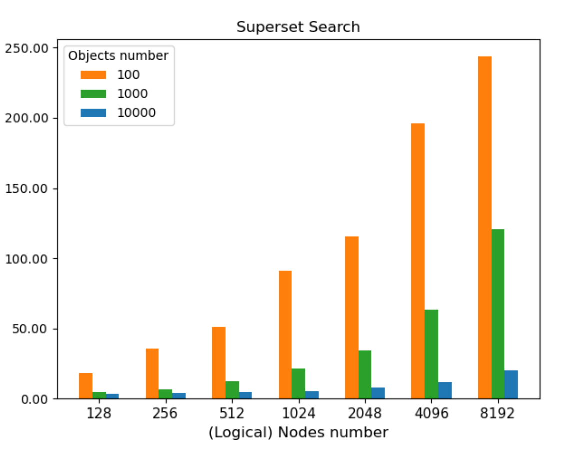
- Average number of hops increases
- from about 3.5 for 128 nodes ($r = 7$)
- to about 6.72 for 8192 nodes ($r = 13$)
- order of the **logarithm of the hypercube logical nodes number** $\rightarrow \frac{\log(n)}{2} = \frac{r}{2}$

Superset Search Results



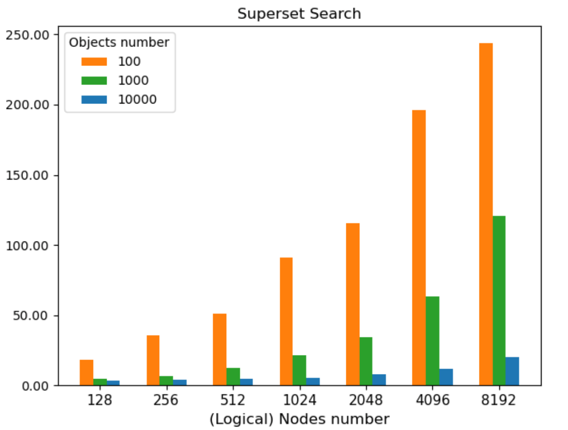
- apparently **anomalous** values stand out

Superset Search Results



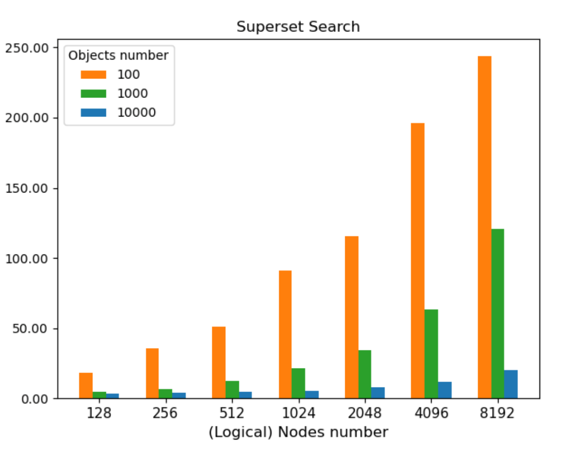
- apparently **anomalous** values stand out
- Superset traverse the network **until it finds the number of objects indicated by the limit**, i.e. $l = 10$

Superset Search Results



- apparently **anomalous** values stand out
- Superset traverse the network **until it finds the number of objects indicated by the limit**, i.e. $l = 10$
- with **many nodes and few objects** $\rightarrow$ the query might take longer to reach that limit, because many nodes are "empty"

Superset Search Results



- apparently **anomalous** values stand out
- Superset traverse the network **until it finds the number of objects indicated by the limit**, i.e. $l = 10$
- with **many nodes and few objects** $\rightarrow$ the query might take longer to reach that limit, because many nodes are "empty"
- order of $\frac{\log(n)}{2} + l$, where l can be set as limit of the nodes number

Conclusion

Conclusion

- **Decentralized data markets** → showing a **DLT layer two solution** → facilitating the retrieval of large amounts of data using **keywords**.

Conclusion

- **Decentralized data markets** $\rightarrow$ showing a **DLT layer two solution** $\rightarrow$ facilitating the retrieval of large amounts of data using **keywords**.
- **IOTA MAM channels** (however,can be easily extended to other DLTs and DFSs).

Conclusion

- **Decentralized data markets** → showing a **DLT layer two solution** → facilitating the retrieval of large amounts of data using **keywords**.
- **IOTA MAM channels** (however,can be easily extended to other DLTs and DFSs).
- The DHT network structured as a hypercube → efficient routing mechanism based on keyword sets.

Conclusion

- **Decentralized data markets** $\rightarrow$ showing a **DLT layer two solution** $\rightarrow$ facilitating the retrieval of large amounts of data using **keywords**.
- **IOTA MAM channels** (however,can be easily extended to other DLTs and DFSs).
- The DHT network structured as a hypercube $\rightarrow$ efficient routing mechanism based on keyword sets.
- Efficient **trade-off between memory space and response time**