

Linking Policies for the Permanent Web

Patrick Le, Quang Bui, Alexander Grigorian, Alexey Kuraev,
Thiyazan Salman, Tu H. Nguyen, Mat Kelly, Sawood Alam

ptl46@drexel.edu jn866@drexel.edu

Drexel University CCI & Internet Archive

ACM/IEEE Joint Conference on Digital Libraries (JCDL)

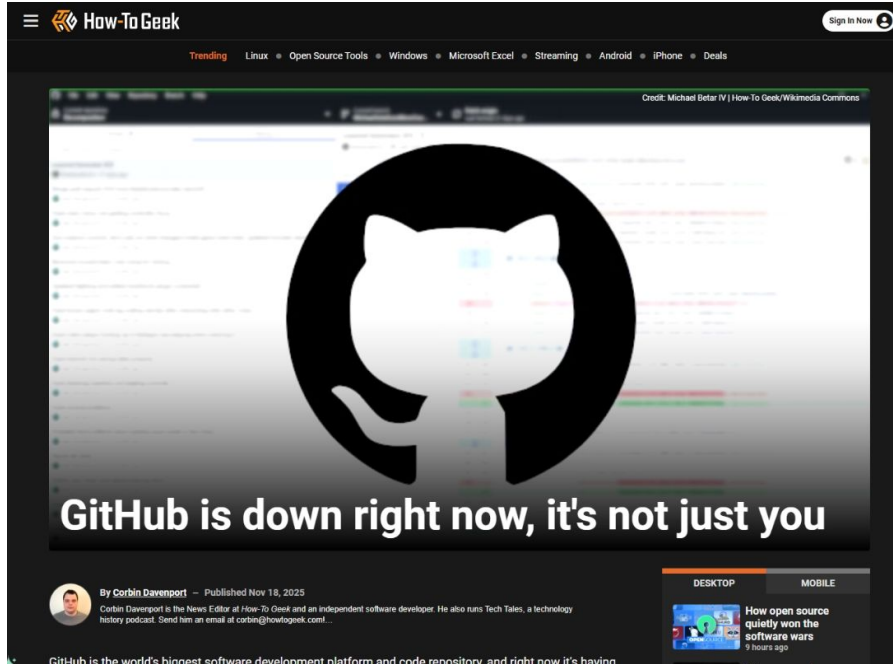
December 18th, 2025



A Senior Capstone Project at Drexel University



Why Decentralize Versioning



<https://www.howtogeek.com/github-is-down-its-not-just-you/>

[+ SECURITY](#) [+ NEWS](#) [+ TECH](#)

The Internet Archive is back as a read-only service after cyberattacks / The Wayback Machine is back online after a data breach and DDoS attacks.

by [+ Tom Warren](#)

Oct 14, 2024, 3:55 PM GMT+7

[Link](#) [Share](#) [38 Comments \(All New\)](#)

INTERNET ARCHIVE
WayBackMachine

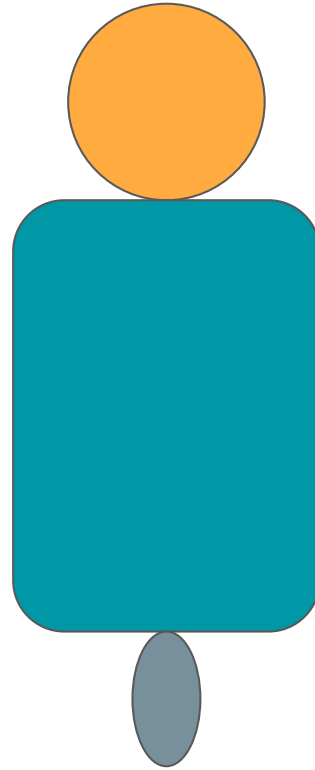
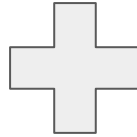
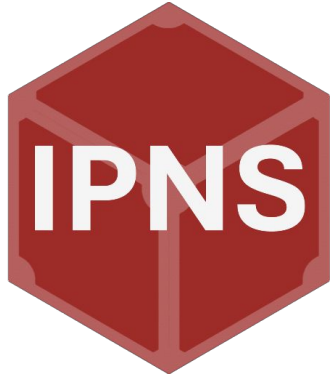
Image: the Internet Archive



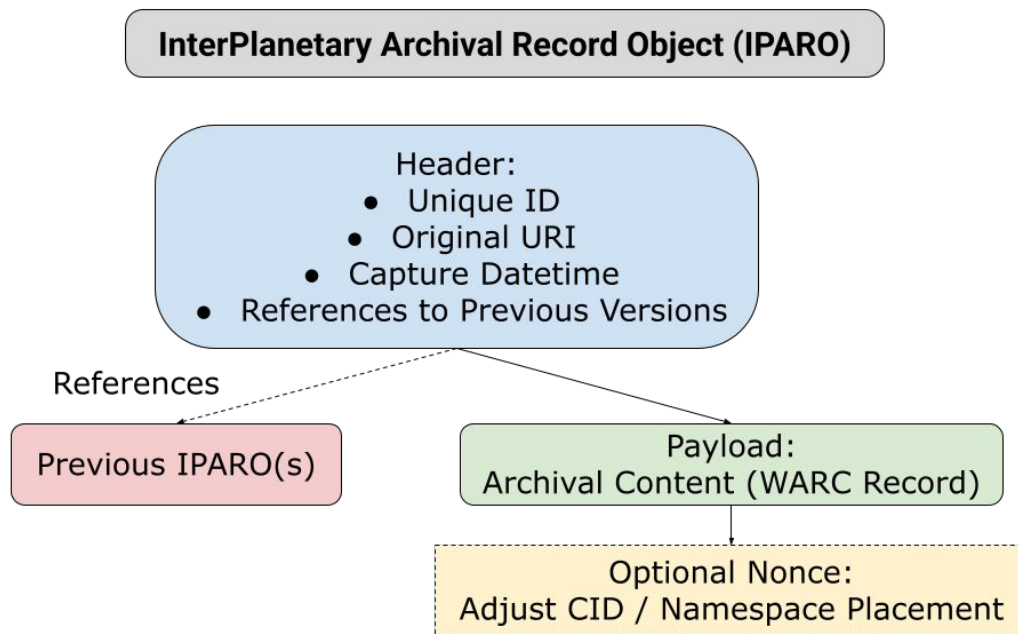
[+ Tom Warren](#) is a senior editor and author of *Notepad*, who has been covering all things Microsoft, PC, and tech for over 20 years.

<https://www.theverge.com/2024/10/14/24269741/internet-archive-online-read-only-data-breach-outage>

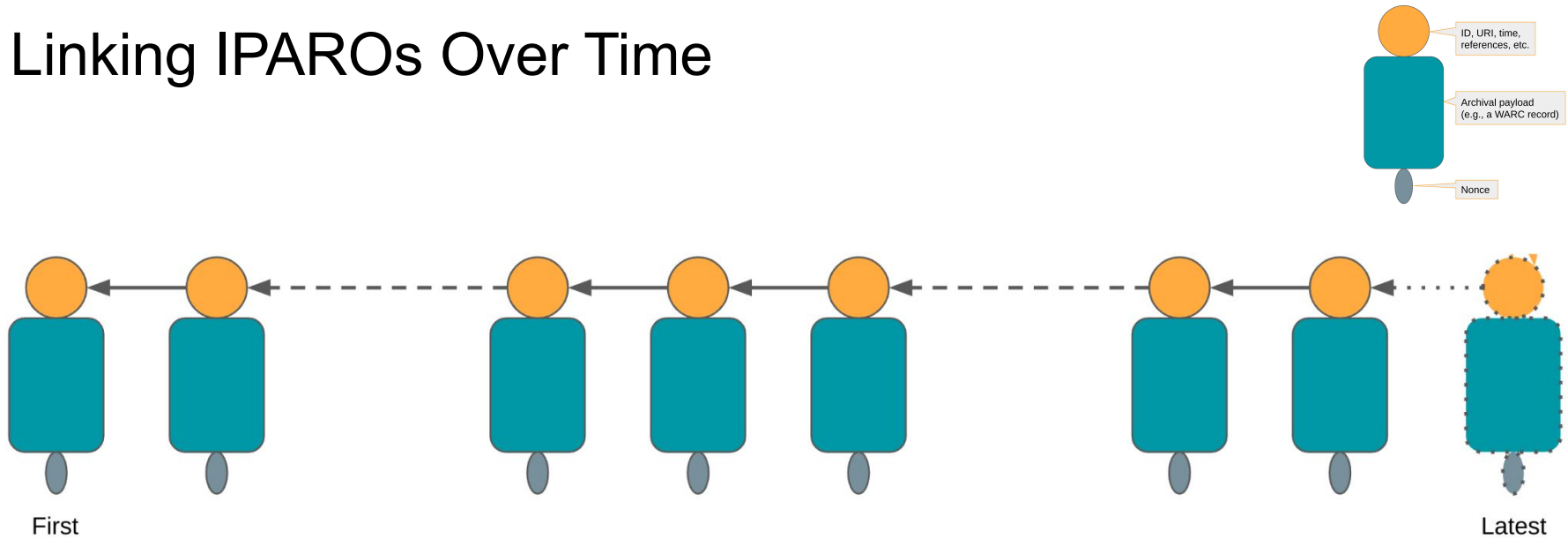
Building Blocks of Decentralized Archiving



What is an IPARO?

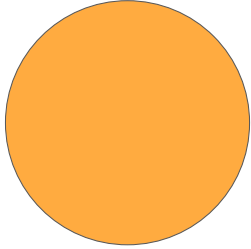


Linking IPAROs Over Time



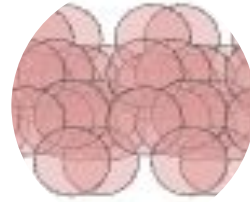
Sawood Alam, "IPARO: InterPlanetary Archival Record Object for Decentralized Web Archiving and Replay," in iPRES 2023.

Methodology



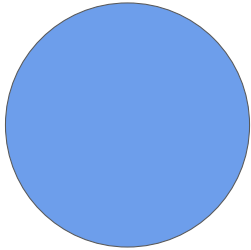
Volume

How many versions for a given object?



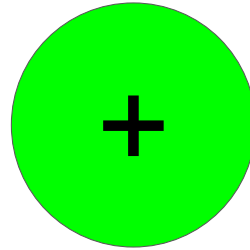
Density

How often are versions recorded?



Linking Strategy





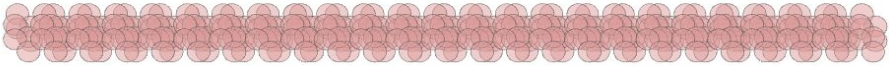
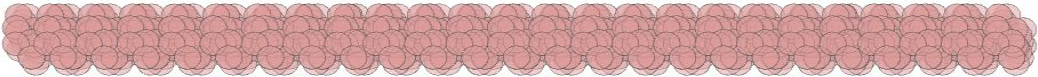
How does each object link to the prior versions?



Operations

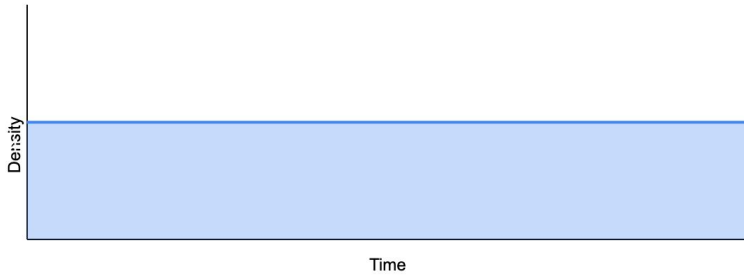
What operations do we perform on these objects?

Version Volumes

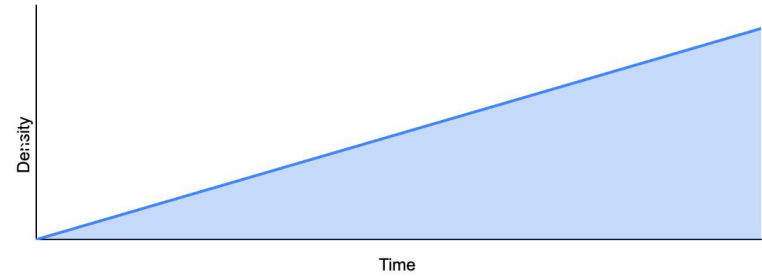
Volume	Range	Example Chain
Single	1	
Tiny	2-9	
Small	10-99	
Medium	100-999	
Large	1000-9999	
Huge	10000+	

Version Densities

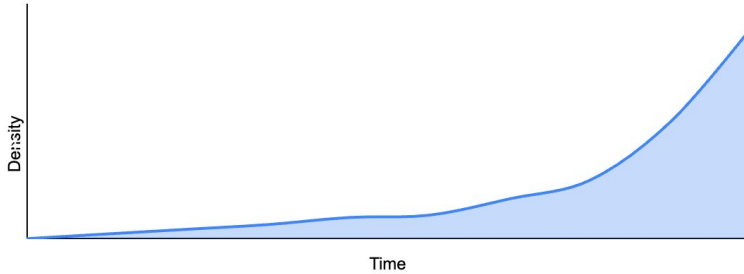
Uniform



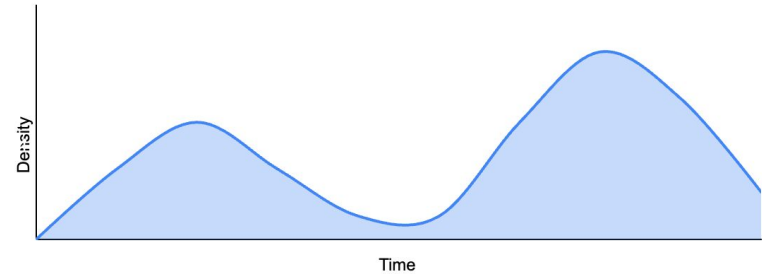
Linear



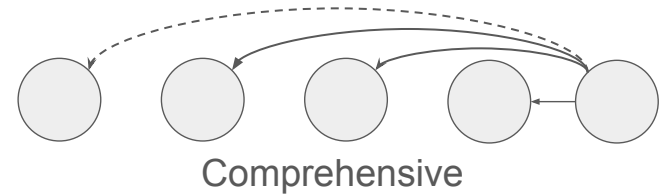
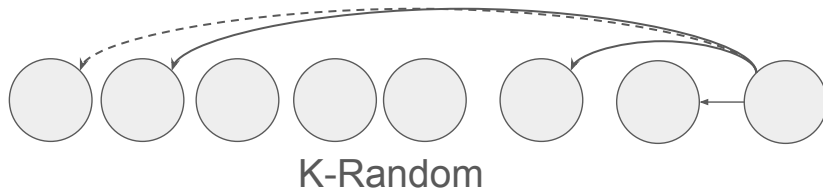
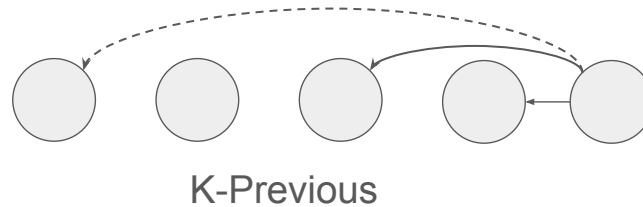
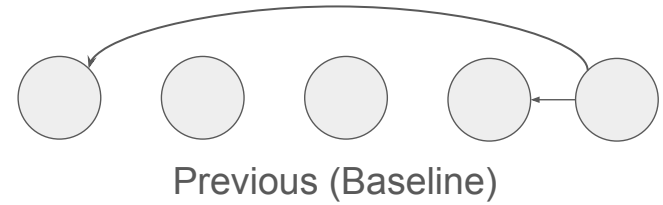
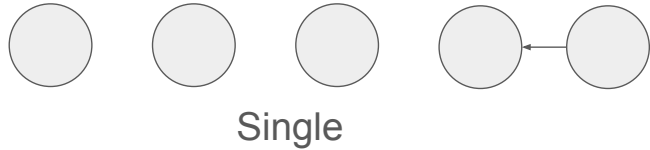
Big Head Long Tail (BHLT)



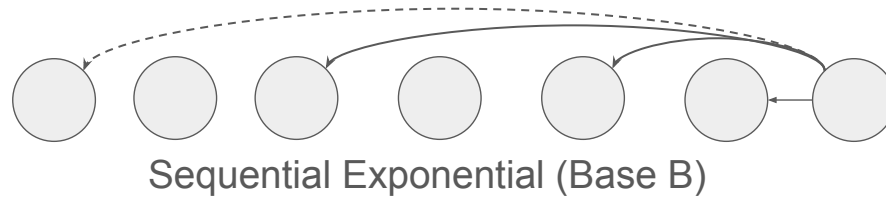
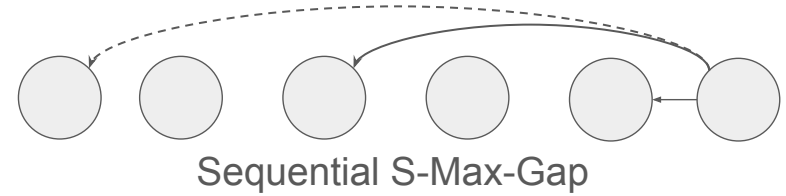
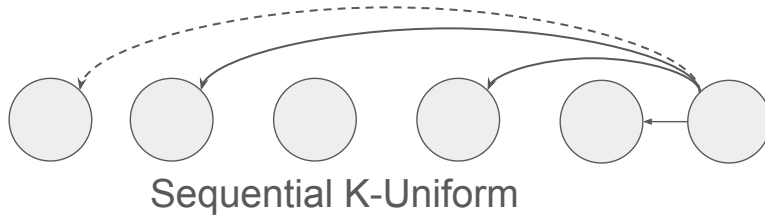
Multipeak



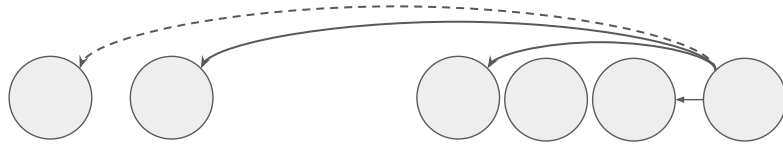
Linking Strategies: Basic



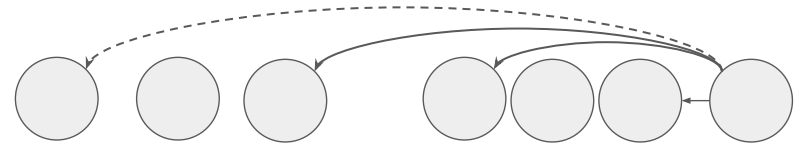
Linking Strategies: Sequential



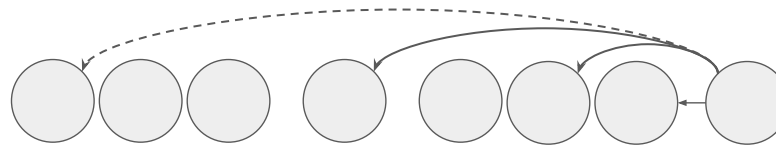
Linking Strategies: Temporal



Temporal K-Uniform



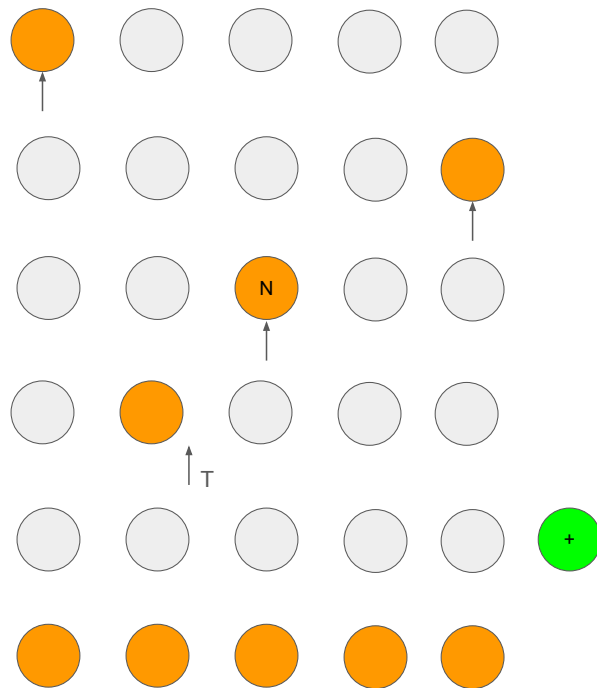
Temporal T-Min-Gap



Temporal Exponential (Base B)

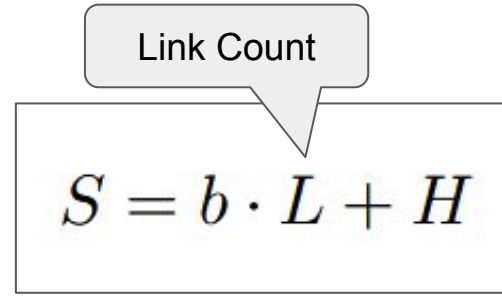
Operations

- Retrieve First
- Retrieve Latest
- Retrieve By Sequence Number (RBSN)
- Retrieve By Time (RBT)
- Add New
- List All



Costs

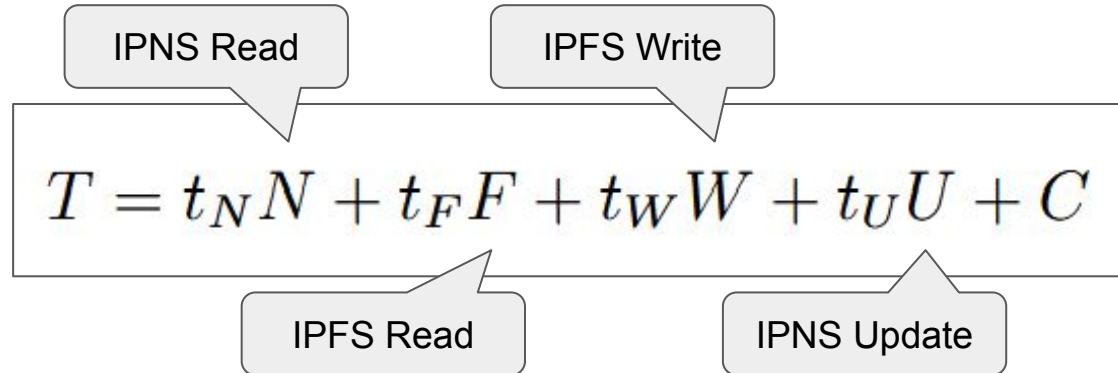
Storage Cost



Link Count

$$S = b \cdot L + H$$

Time Cost



IPNS Read

IPFS Write

$$T = t_N N + t_F F + t_W W + t_U U + C$$

IPFS Read

IPNS Update

Complexities

Strategy	Links Per Node	Add Node	RBT	RBSN	List All
Single	1	1	N	N	N
Previous	1	1	N	N	N
Comprehensive	N	1	1	1	1
K -Previous	K	K	N'	N'	N'
K -Random	K	\sqrt{NK}	$\sqrt{N'}$	$\sqrt{N'}$	N'
Sequential K -Uniform	K	K	N'	N'	N'
Sequential S -Max-Gap	N/S	N/S	S	S	S
Sequential Base B Exponential	$\log N$	$\log^2 N$	$\log N$	$\log N$	N
Temporal K -Uniform	K	N'	$N_{T'}$	$N_{T'}$	N
Temporal T -Min-Gap	\bar{T}	\bar{T}	$\bar{T} + N_T$	$\bar{T} + N_T$	N
Temp-Exp Base B (Time Unit T)	$\log(\bar{T})$	$N_T \log \bar{T}$	N_T	N_T	$N_T \log \bar{T}$

- $N' = N/K$
- T^* measures time between first and latest node
- $T' = T^*/K$
- $\bar{T} = T^*/T$
- N_T is the maximum nodes for time window T .

Complexities

Strategy	Links Per Node	Add Node	RBT	RBSN	List All
Single	1	1	N	N	N
Previous	1	1	N	N	N
Comprehensive	N	1	1	1	1
K -Previous	K	K	N'	N'	N'
K -Random	K	\sqrt{NK}	$\sqrt{N'}$	$\sqrt{N'}$	N'
Sequential K -Uniform	K	K	N'	N'	N'
Sequential S -Max-Gap	N/S	N/S	S	S	S
Sequential Base B Exponential	$\log N$	$\log^2 N$	$\log N$	$\log N$	N
Temporal K -Uniform	K	N'	$N_{T'}$	$N_{T'}$	N
Temporal T -Min-Gap	\bar{T}	\bar{T}	$\bar{T} + N_T$	$\bar{T} + N_T$	N
Temp-Exp Base B (Time Unit T)	$\log(\bar{T})$	$N_T \log \bar{T}$	N_T	N_T	$N_T \log \bar{T}$

- $N' = N/K$
- T^* measures time between first and latest node
- $T' = T^*/K$
- $\bar{T} = T^*/T$
- N_T is the maximum nodes for time window T .

Complexities

Strategy	Links Per Node	Add Node	RBT	RBSN	List All
Single	1	1	N	N	N
Previous	1	1	N	N	N
Comprehensive	N	1	1	1	1
K -Previous	K	K	N'	N'	N'
K -Random	K	\sqrt{NK}	$\sqrt{N'}$	$\sqrt{N'}$	N'
Sequential K -Uniform	K	K	N'	N'	N'
Sequential S -Max-Gap	N/S	N/S	S	S	S
Sequential Base B Exponential	$\log N$	$\log^2 N$	$\log N$	$\log N$	N
Temporal K -Uniform	K	N'	$N_{T'}$	$N_{T'}$	N
Temporal T -Min-Gap	\bar{T}	\bar{T}	$\bar{T} + N_T$	$\bar{T} + N_T$	N
Temp-Exp Base B (Time Unit T)	$\log(\bar{T})$	$N_T \log \bar{T}$	N_T	N_T	$N_T \log \bar{T}$

- $N' = N/K$
- T^* measures time between first and latest node
- $T' = T^*/K$
- $\bar{T} = T^*/T$
- N_T is the maximum nodes for time window T .

Experimental System

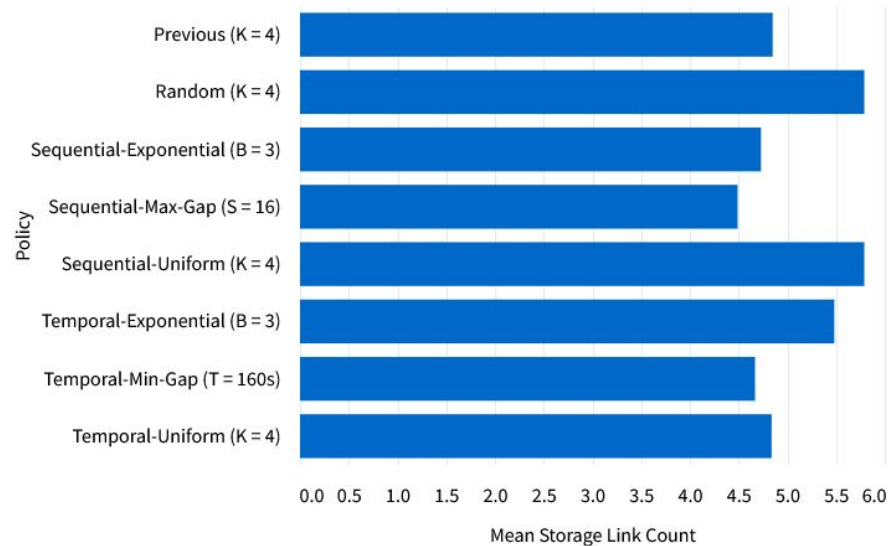
The screenshot shows a web browser window with the address bar displaying `http://localhost:8501/General_Settings`. The page title is "General Settings". On the left, there is a sidebar menu with the following sections: "Home" (with a link to Home), "Settings" (with "General Settings" selected and a link to "Select Policies"), "Reports" (with links to "Summary Report", "Space Time Tradeoff", "Policy Growth Rate", "Cost Map", "Iteration Level Analysis", "Cost Density Map", "Cost Calculator", "Resilience Report", and "Policy Growth Rate Heatmap"), "Visualization" (with links to "Policy Visualization Settings" and "Policy Visualization"), and "Simulation (WIP)" (with links to "Add Policies", "View Environments", "Add Version Densities", and "Simulation Writer Output"). The main content area is titled "General Settings" and includes a note: "You must fill in the general settings before selecting policies and accessing reports." Below this note is a form with three sections: "Filter Settings" (containing "Version Density Settings" with a "Density" dropdown menu set to "Uniform"), "Version Volume Settings" (with a "Version Volume" dropdown menu set to "1"), and "Graph Display Settings" (with a "Logarithmic Scale" checkbox and a "Submit" button). The browser's address bar shows a zoom level of 80% and a "Sign in" button.

<https://github.com/johnnguyenn77/iparo>

Experimental Setup

- Windows 11 PC, 8 cores
- Chains have length 100
- 10 iterations per operation

Chosen Strategies



Experimental Storage and Time Cost

	Single	Comprehensive	Exponential
Retrieve First	100	2	2
Retrieve Latest	1	1	1
Retrieve By Seq. #	54	2	4.5
Retrieve By Time	54.4	2	4
Add New	0.99	0.99	6.65
List All	99	1	49
Links Per Node	0.99	49.5	6.65

Nodes: 100, Density: Uniform, Exp. Base: 2

Experimental Storage and Time Cost

	Single	Comprehensive	Exponential
Retrieve First	100	2	2
Retrieve Latest	1	1	1
Retrieve By Seq. #	54	2	4.5
Retrieve By Time	54.4	2	4
Add New	0.99	0.99	6.65
List All	99	1	49
Links Per Node	0.99	49.5	6.65

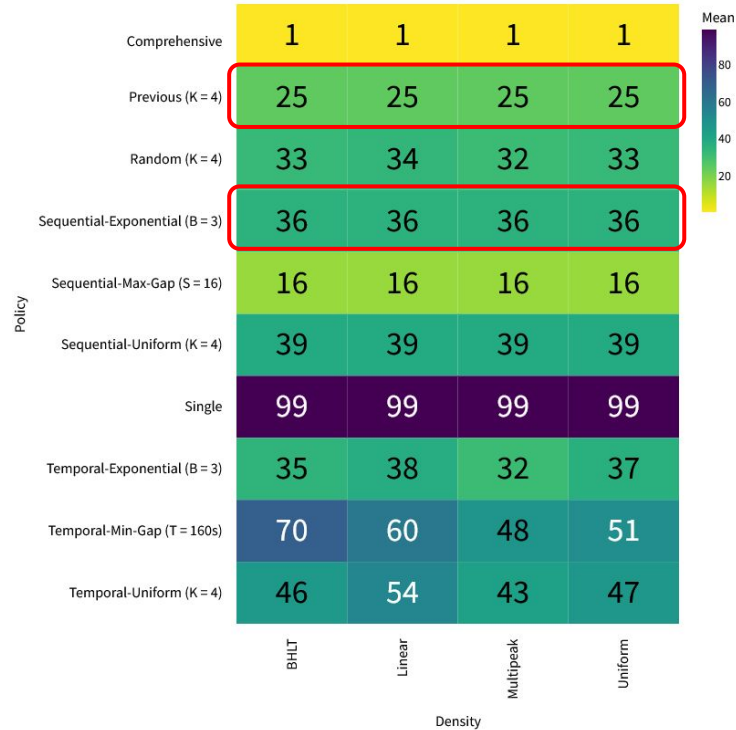
Nodes: 100, Density: Uniform, Exp. Base: 2

Experimental Storage and Time Cost

	Single	Comprehensive	Exponential
Retrieve First	100	2	2
Retrieve Latest	1	1	1
Retrieve By Seq. #	54	2	4.5
Retrieve By Time	54.4	2	4
Add New	0.99	0.99	6.65
List All	99	1	49
Links Per Node	0.99	49.5	6.65

Nodes: 100, Density: Uniform, Exp. Base: 2

Time Cost: List All



Conclusion and Future Work

- Which strategies to use?
 - Exponential: General Purpose
 - K-Previous: Many List All operations
- Future Work
 - Sampling other distributions
 - Which strategies are the most resilient?

Summary

Decentralized Versioning

IPFS
IPNS
IPARO

System Configurations

6 Volumes
4 Densities
11 Strategies
6 Operations
4 Parameters

Cost Analysis

Storage (for links)
Time (for operations)
Theoretical Complexities
Experimental Results

<https://github.com/johnnguyenn77/iparo>