A Collaborative, Secure, and Private InterPlanetary Wayback Web Archiving System Using IPFS

Mat Kelly
Old Dominion University
Norfolk, Virginia, USA
@machawk1

David Dias
Protocol Labs
Planet Earth
@daviddias

https://github.com/oduwsdl/ipwb

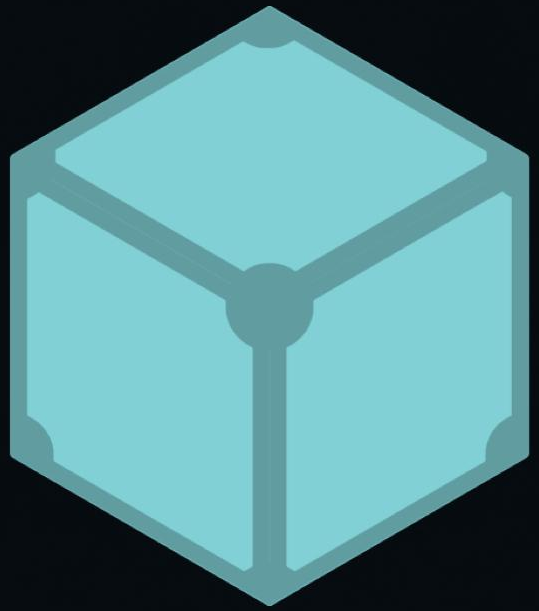
https://ipfs.io

w/ Sawood Alam, Michael L. Nelson, and Michele C. Weigle

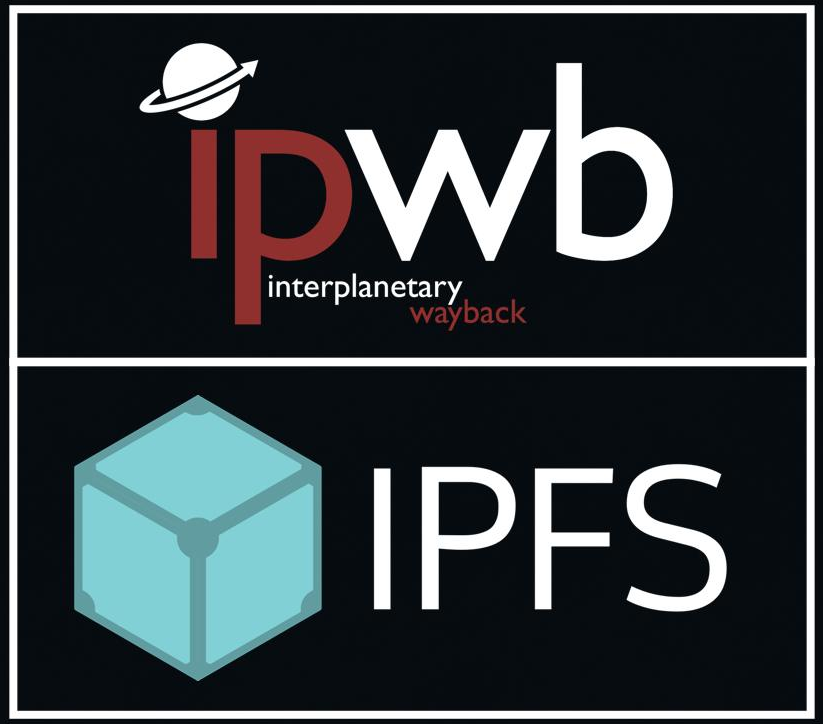
IIPC Web Archiving Conference • June 15, 2017 • London, UK

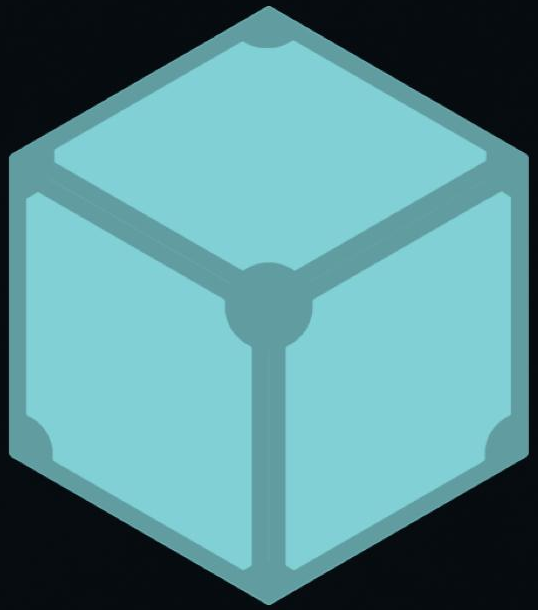
Outline

- InterPlanetary File System Motivation & Design
- InterPlanetary Wayback Motivation & Design
- How IPFS/IPWB relate, relevancy to Web archiving
- Advances in IPFS/IPWB
- Demo(s)



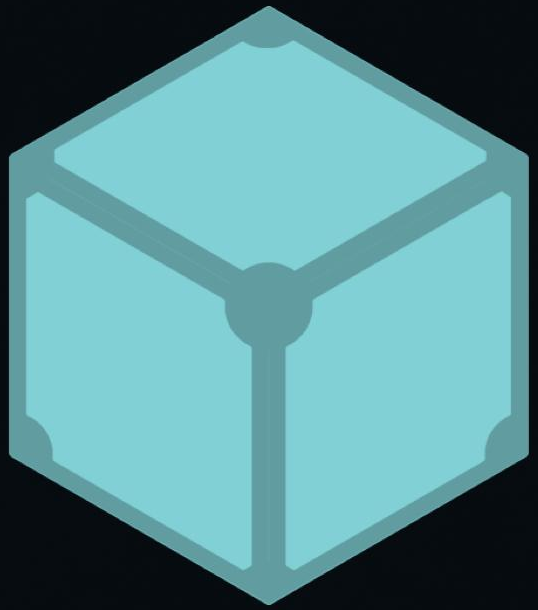
IPFS


ipwb
interplanetary
wayback
IPFS



IPFS

InterPlanetary FileSystem



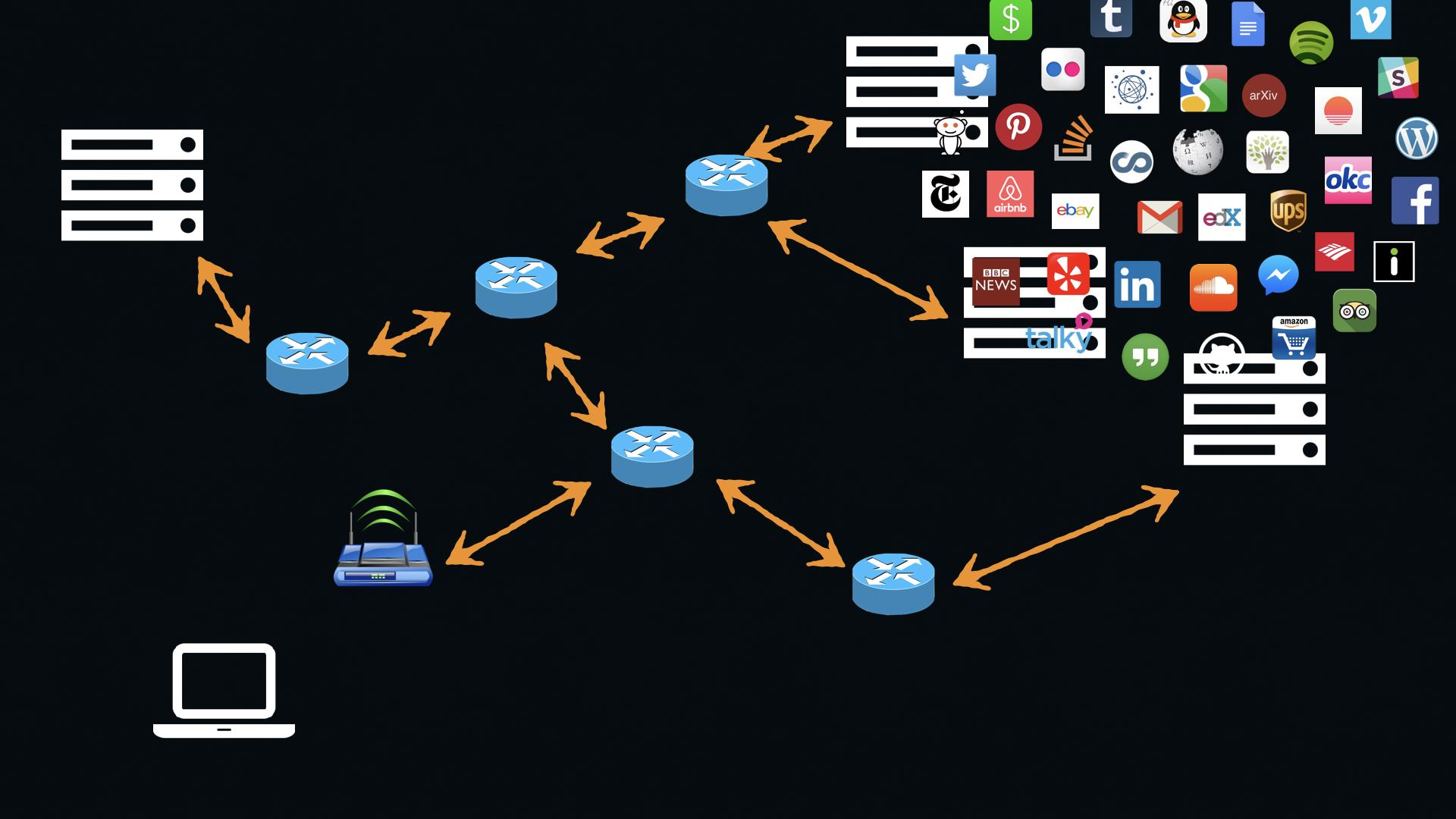
IPFS

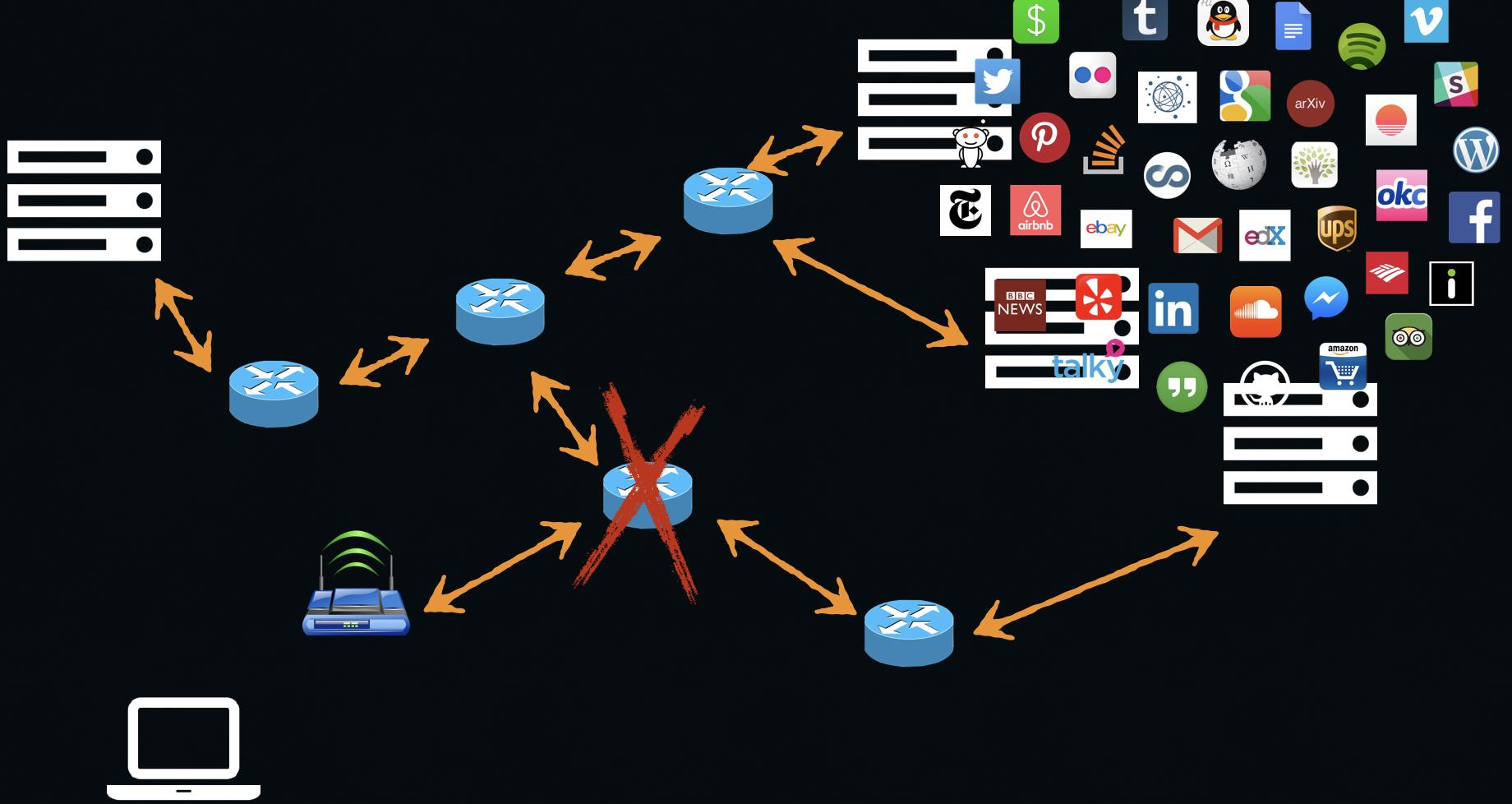
InterPlanetary FileSystem

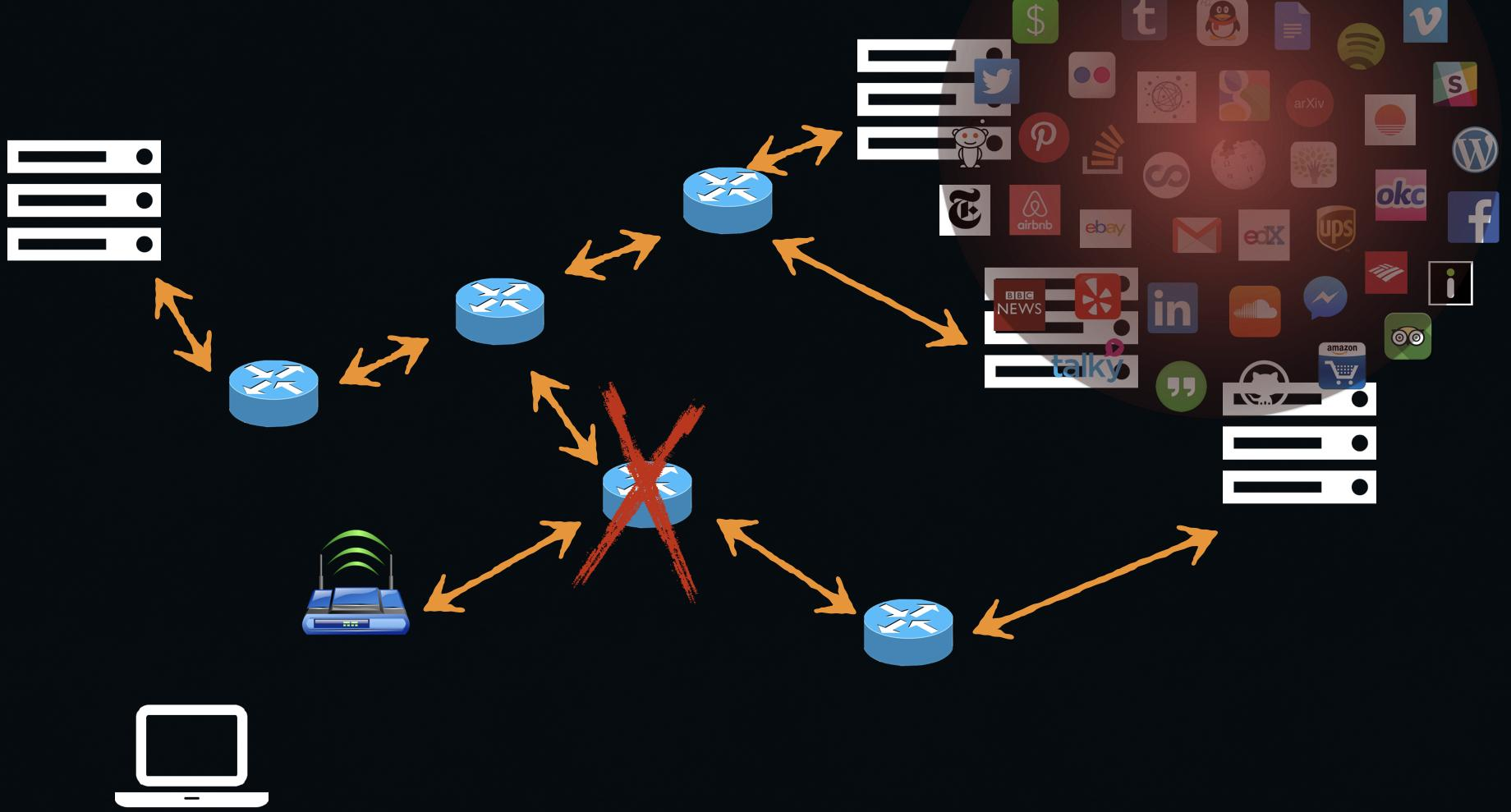
Location → Content Addressing


www

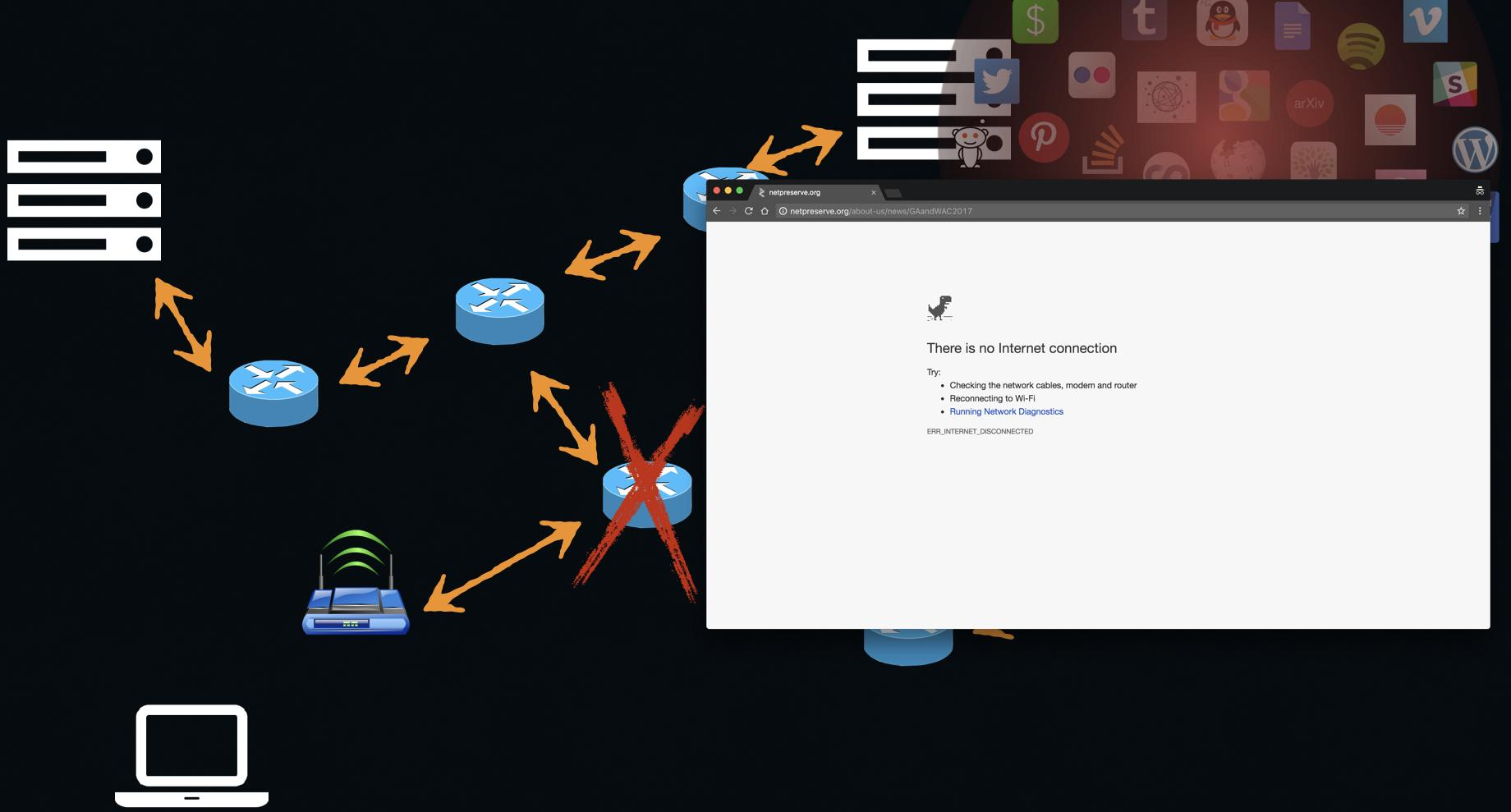
Offline Capabilities


BBC NEWS
talky
airbnb
ebay
arXiv
okc
edX
ups
amazon


BBC NEWS
talky
airbnb
ebay
arXiv
okc
edX
amazon


BBC NEWS
talky
airbnb
ebay
edX
ups
okc
amazon
arXiv


ABOUT IIPC-draft | IIPC
netpreserve.org/about-us/news/GAandWAC2017
IIPC
netpreserve.org
INTERNATIONAL INTERNET PRESERVATION CONSORTIUM
HOME
ABOUT IIPC
WEB ARCHIVING
EVENTS
BLOG
JOIN US
ABOUT IIPC-DRAFT
Who is the IIPC?
IIPC members have the unique expertise to collect, preserve and make accessible knowledge from the global web.
International Internet Preservation Consortium
About the IIPC
INTERNATIONAL INTERNET PRESERVATION CONSORTIUM
HISTORY
In July 2003, the IIPC was formally chartered at the National Library of France with 12 participating institution.
The initial agreement was in effect for three years, and membership was limited to charter institutions. The IIPC is now open to libraries, archives,


netpreserve.org
netpreserve.org/about-us/news/GAandWAC2017
There is no Internet connection
Try:
Checking the network cables, modem and router
Reconnecting to Wi-Fi
Running Network Diagnostics
ERR_INTERNET_DISCONNECTED

Permanence

http://location.site/important-data

Protocol

Location

Content path

~~http://location.site/important-data~~

Protocol

Location

Content path

http://other.site/important-data

~~http://location.site/important-data~~

Protocol

Location

Content path

http://other.site/important-data

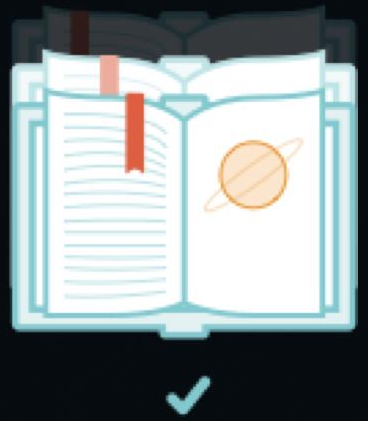
~~http://location.site/important-data~~

Protocol

Location

Content path

http://other.site/important-data



Content Addressing

IP:120.1.11.22

IP:15.35.32.21

IP:12.1.11.22

IP:10.20.30.40

http://example.com/cat.png

http://10.20.30.40/cat.png

location

http://example.com/cat.png

↓

http://10.20.30.40/cat.png

location

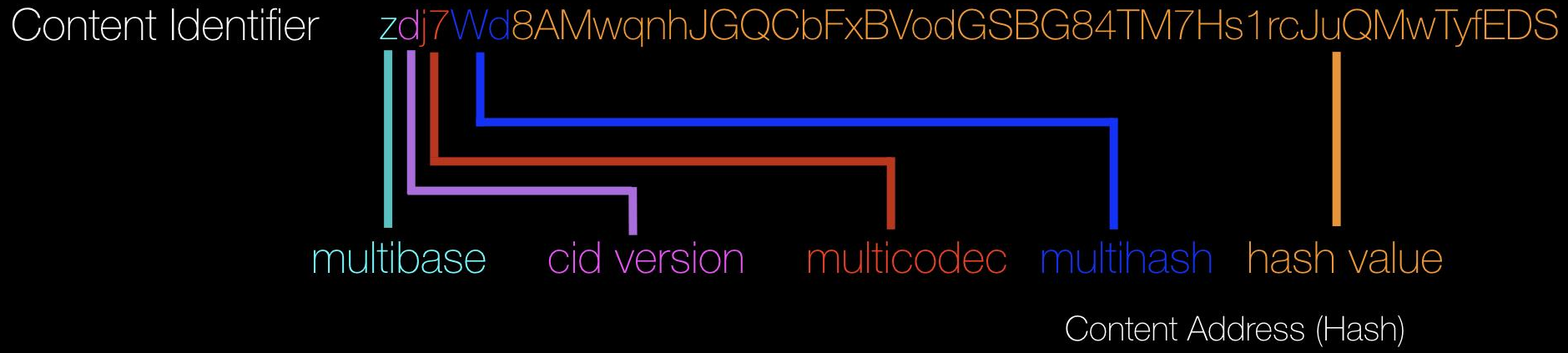
/ipns/example.com/cat.png

↓

/ipfs/QmW98pJrc6FZ6/cat.png

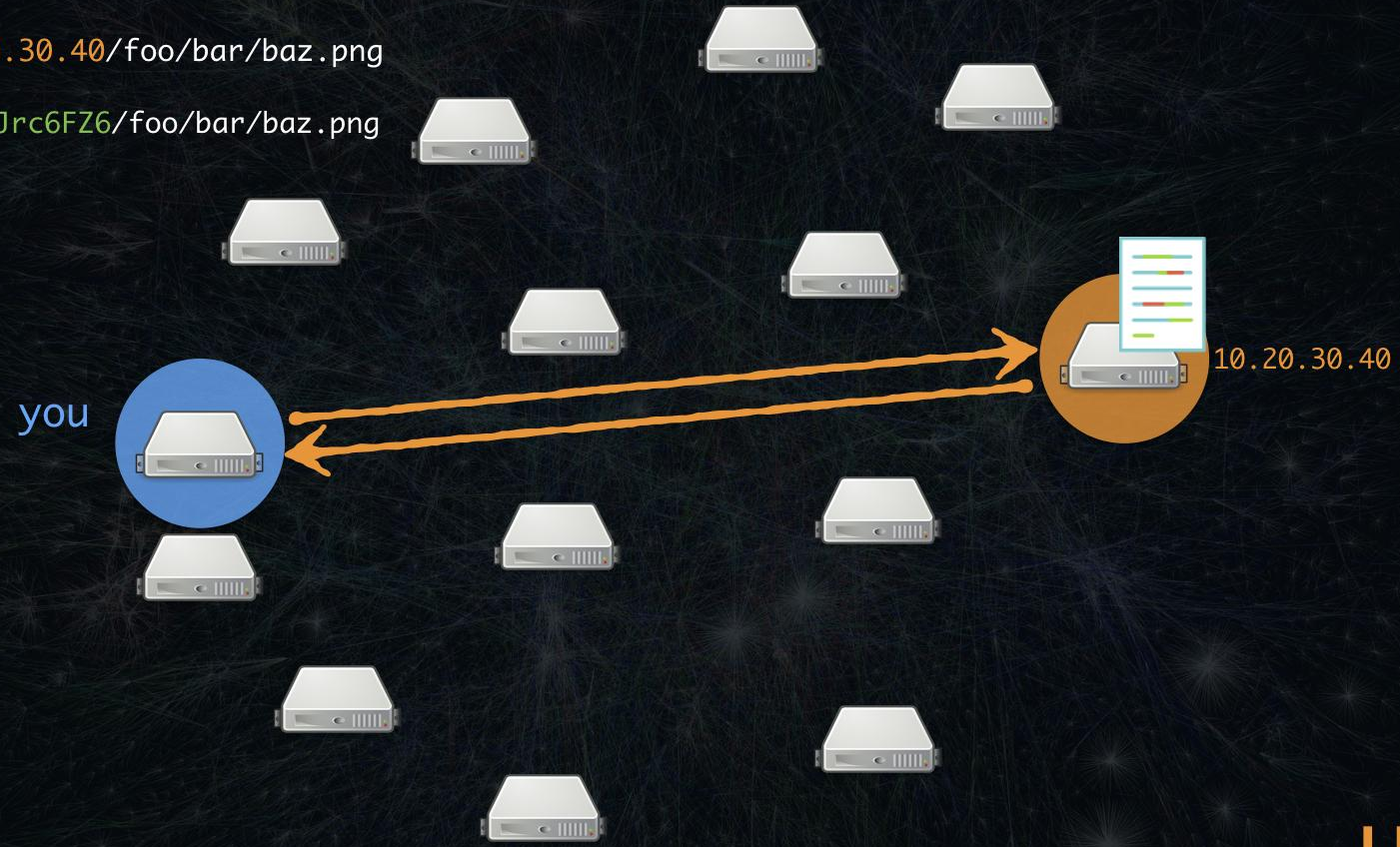
content

CID - Content Identifier



http://10.20.30.40/foo/bar/baz.png

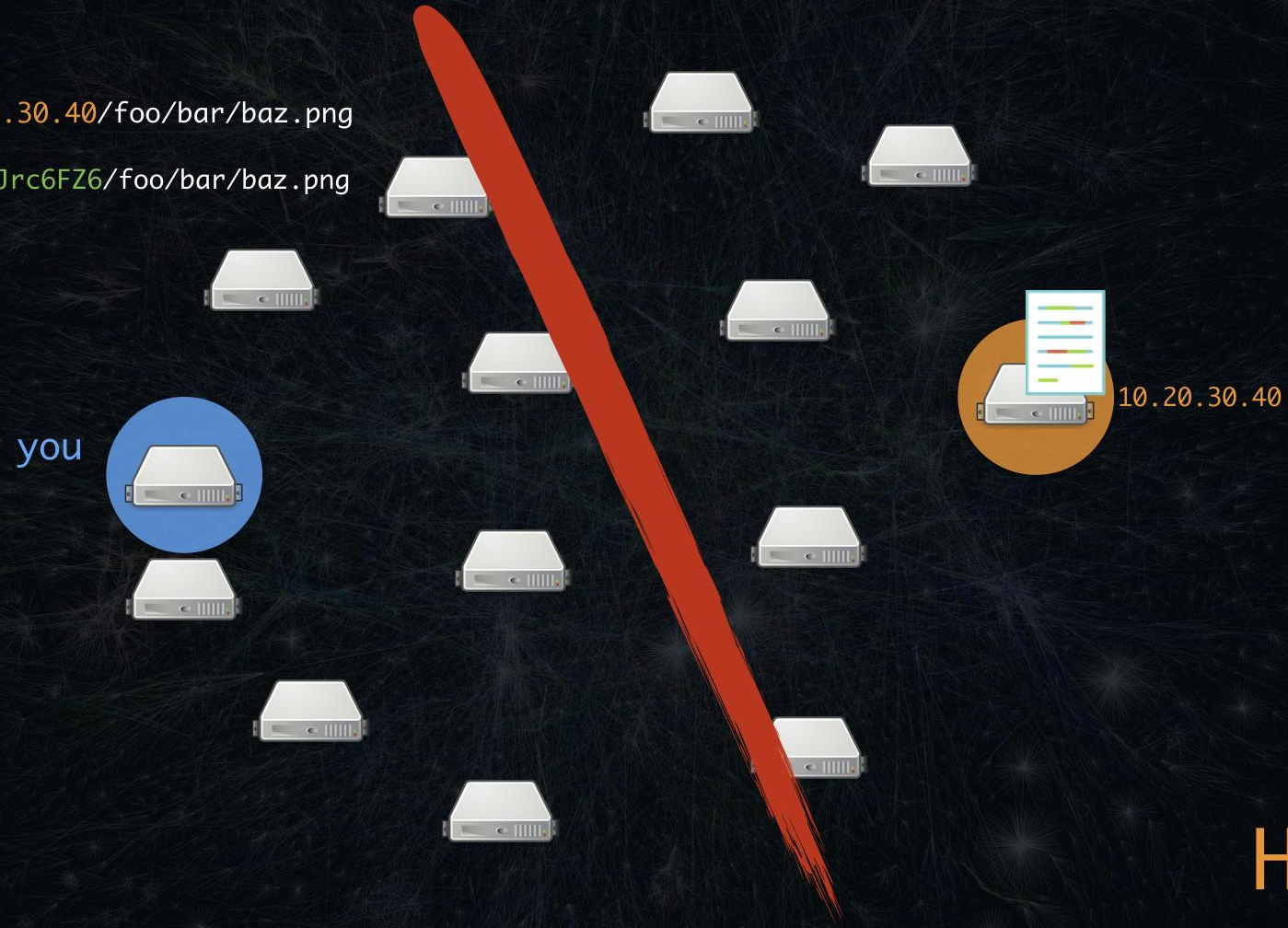
/ipfs/QmW98pJrc6FZ6/foo/bar/baz.png



HTTP

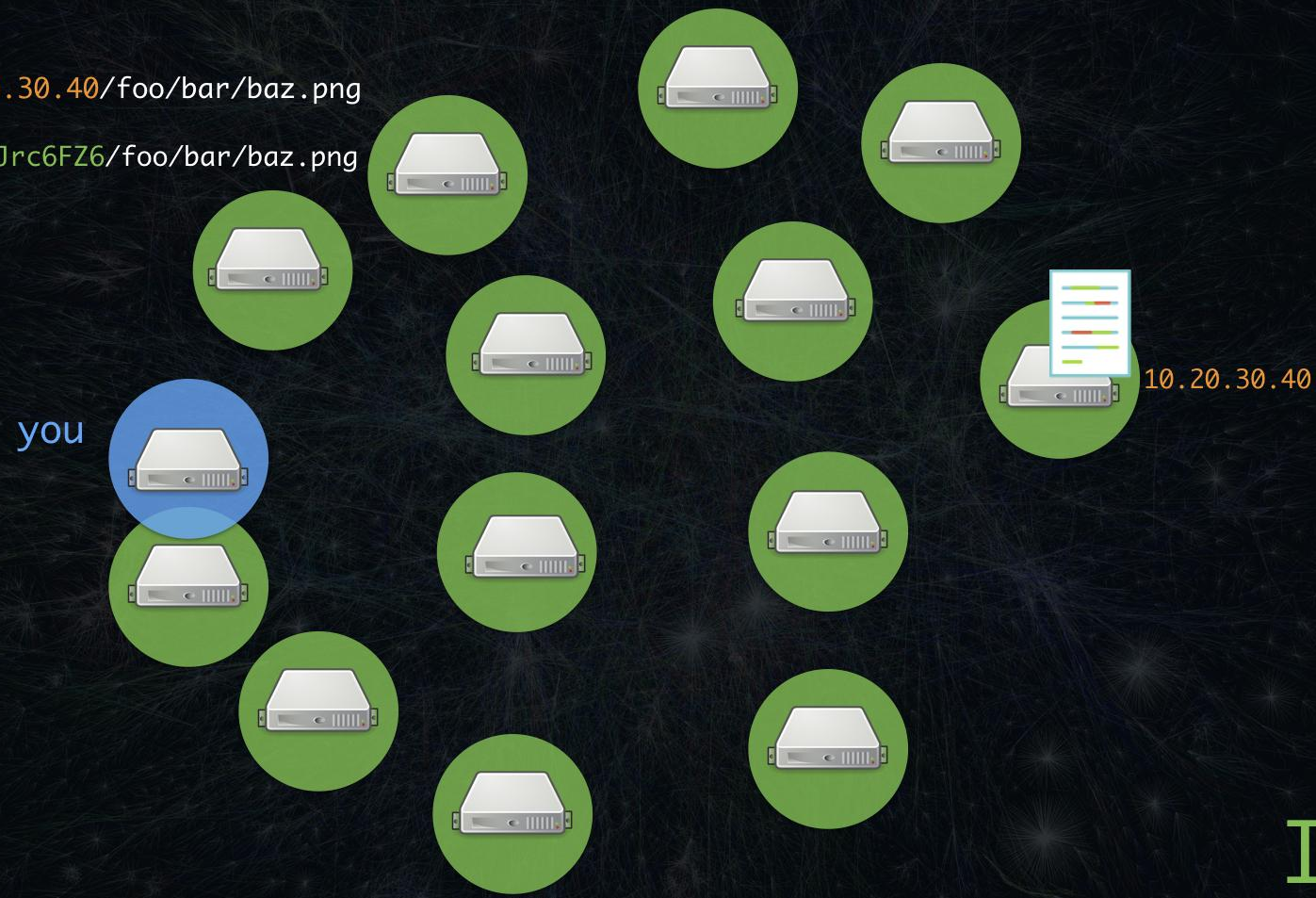
http://10.20.30.40/foo/bar/baz.png

/ipfs/QmW98pJrc6FZ6/foo/bar/baz.png



http://10.20.30.40/foo/bar/baz.png

/ipfs/QmW98pJrc6FZ6/foo/bar/baz.png



IPFS

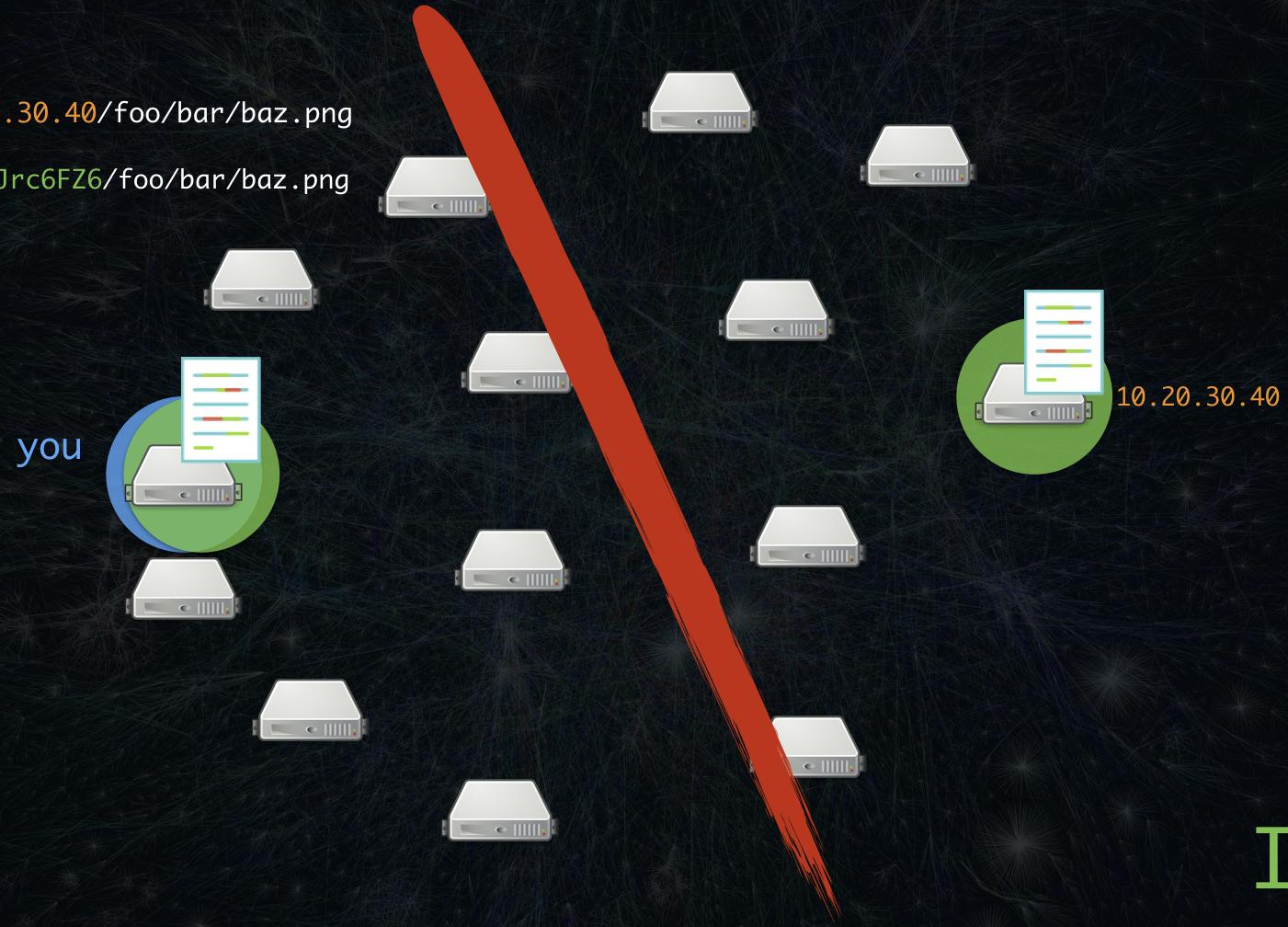
http://10.20.30.40/foo/bar/baz.png

/ipfs/QmW98pJrc6FZ6/foo/bar/baz.png

IPFS

http://10.20.30.40/foo/bar/baz.png

/ipfs/QmW98pJrc6FZ6/foo/bar/baz.png



Disconnected

Control

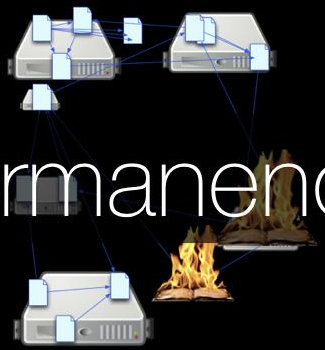
200 MB x 30 x 8 = 48 GB

Bandwidth

IoT

Offline

Permanence



Security

find out more

Epicenter Bitcoin Interview

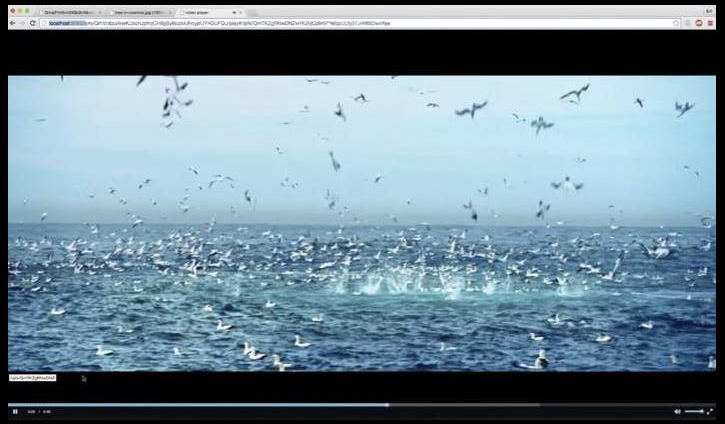
youtu.be/erB7i6Uc4DM

IPFS Talk at Stanford

youtu.be/HUVmypx9HGI

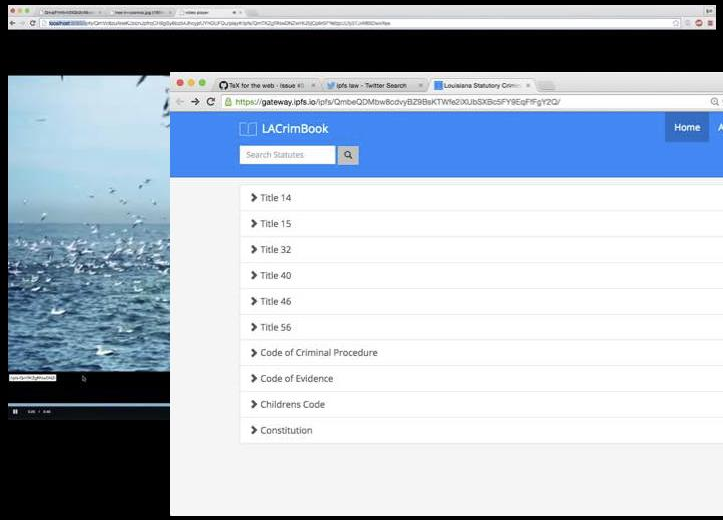
Join us on GitHub!

github.com/ipfs/ipfs



video distribution + streaming

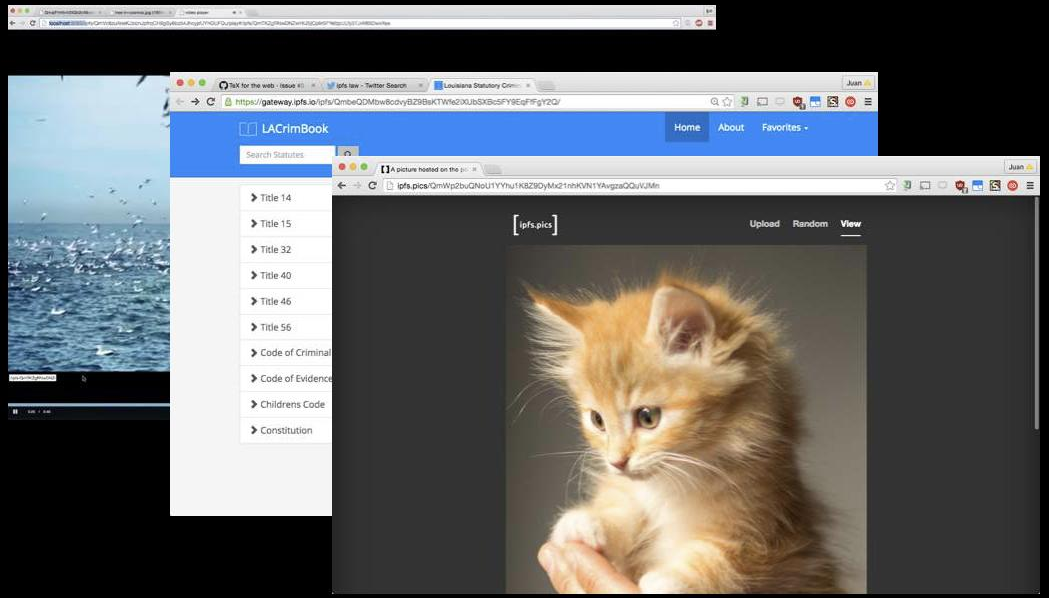
Live Examples



legal documents

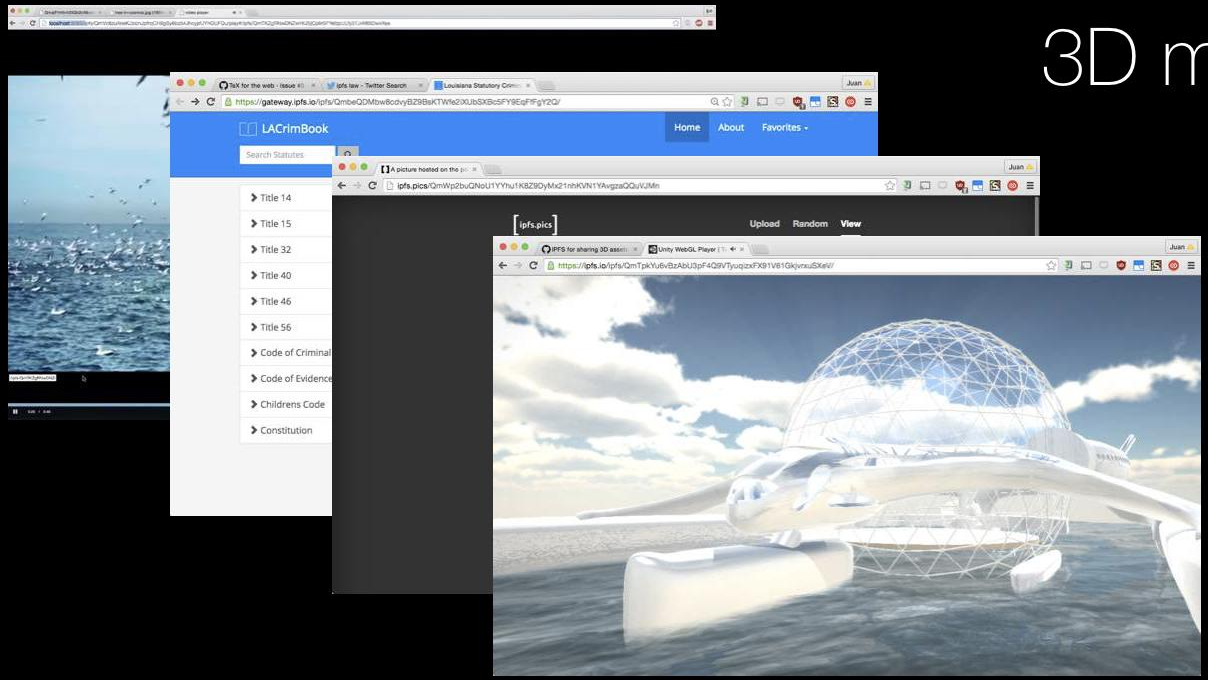
Live Examples

ipfs.pics (imgur-like)



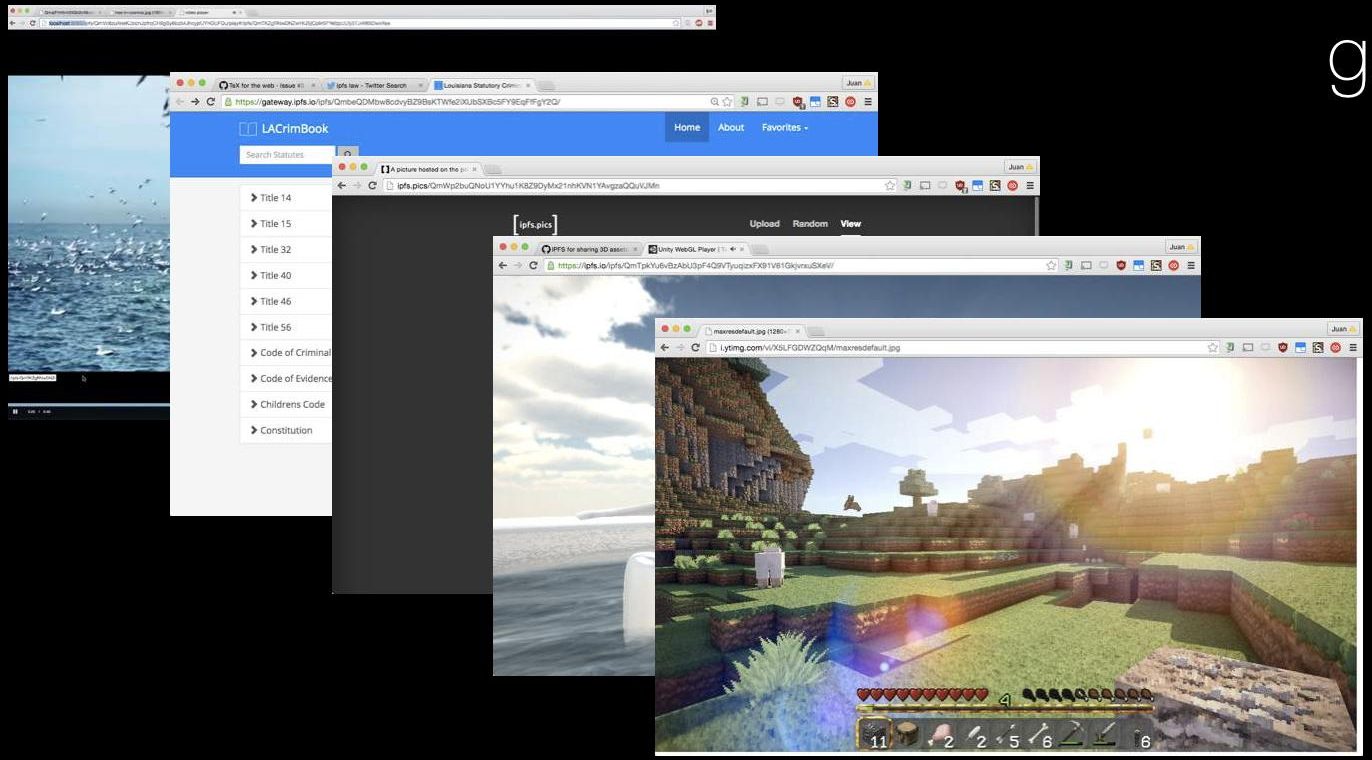
Live Examples

3D models (they're big!)



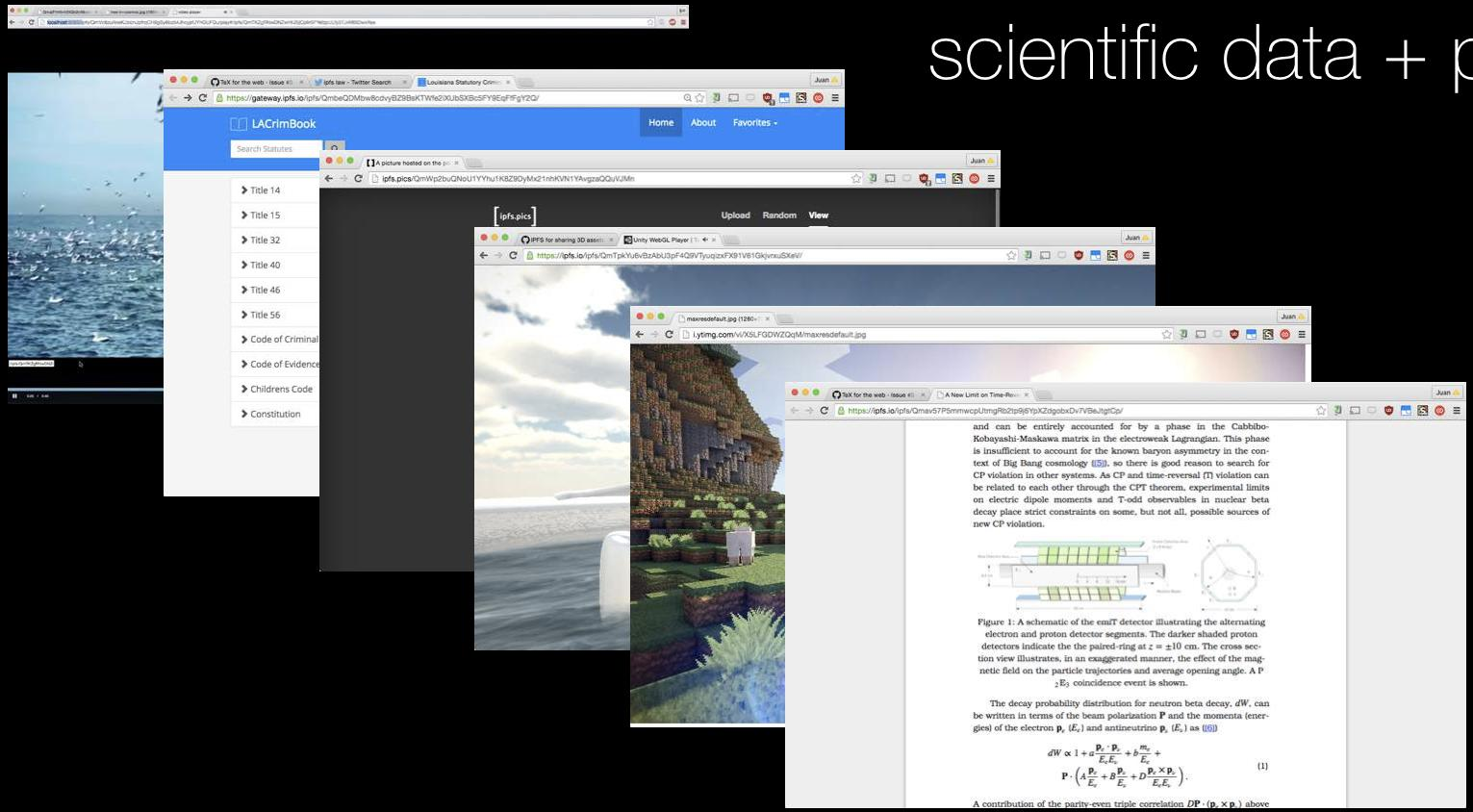
Live Examples

games



Live Examples

scientific data + papers



and can be entirely accounted for by a phase in the Cabbibo-Kobayashi-Maskawa matrix in the electroweak Lagrangian. This phase is insufficient to account for the known baryon asymmetry in the context of Big Bang cosmology ([5]), so there is good reason to search for CP violation in other systems. As CP and time-reversal [T] violation can be related to each other through the CPT theorem, experimental limits on electric dipole moments and T-odd observables in nuclear beta decay place strict constraints on some, but not all, possible sources of new CP violation.

Figure 1: A schematic of the emiT detector illustrating the alternating electron and proton detector segments. The darker shaded proton detectors indicate the the paired-ring at $z = \pm 10$ cm. The cross section view illustrates, in an exaggerated manner, the effect of the magnetic field on the particle trajectories and average opening angle. A P $_2$E$_3$ coincidence event is shown.

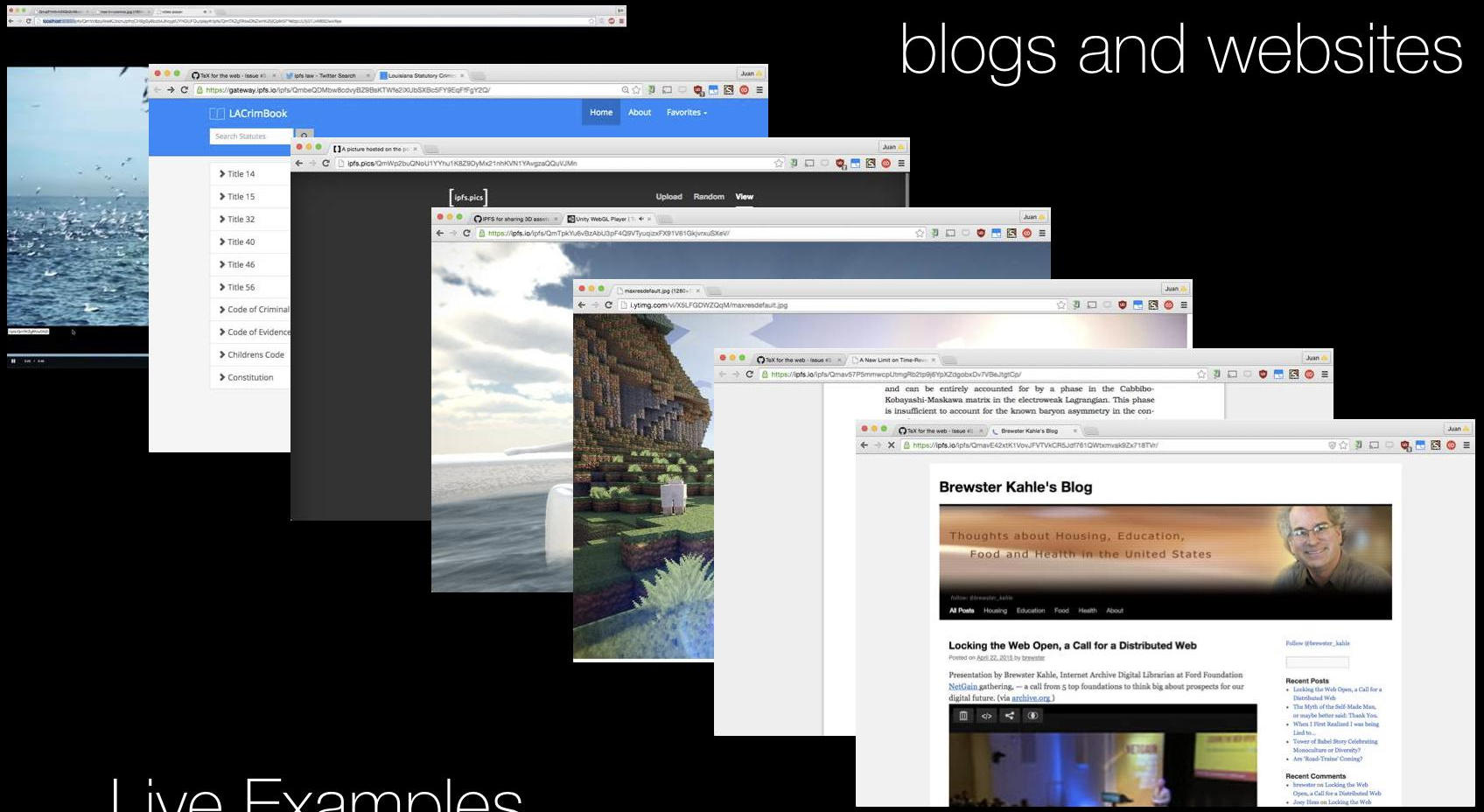
The decay probability distribution for neutron beta decay, dW, can be written in terms of the beam polarization $\mathbf{P}$ and the momenta (energies) of the electron $\mathbf{p}_e$ (E_e) and antineutrino $\mathbf{p}_\nu$ (E_ν) as ([6])

$$dW \propto 1 + a\frac{\mathbf{p}_e \cdot \mathbf{p}_\nu}{E_e E_\nu} + b\frac{m_e}{E_e} + \mathbf{P} \cdot \left(A\frac{\mathbf{p}_e}{E_e} + B\frac{\mathbf{p}_\nu}{E_\nu} + D\frac{\mathbf{p}_e \times \mathbf{p}_\nu}{E_e E_\nu} \right). \quad (1)$$

A contribution of the parity-even triple correlation $D\mathbf{P} \cdot (\mathbf{p}_e \times \mathbf{p}_\nu)$ above

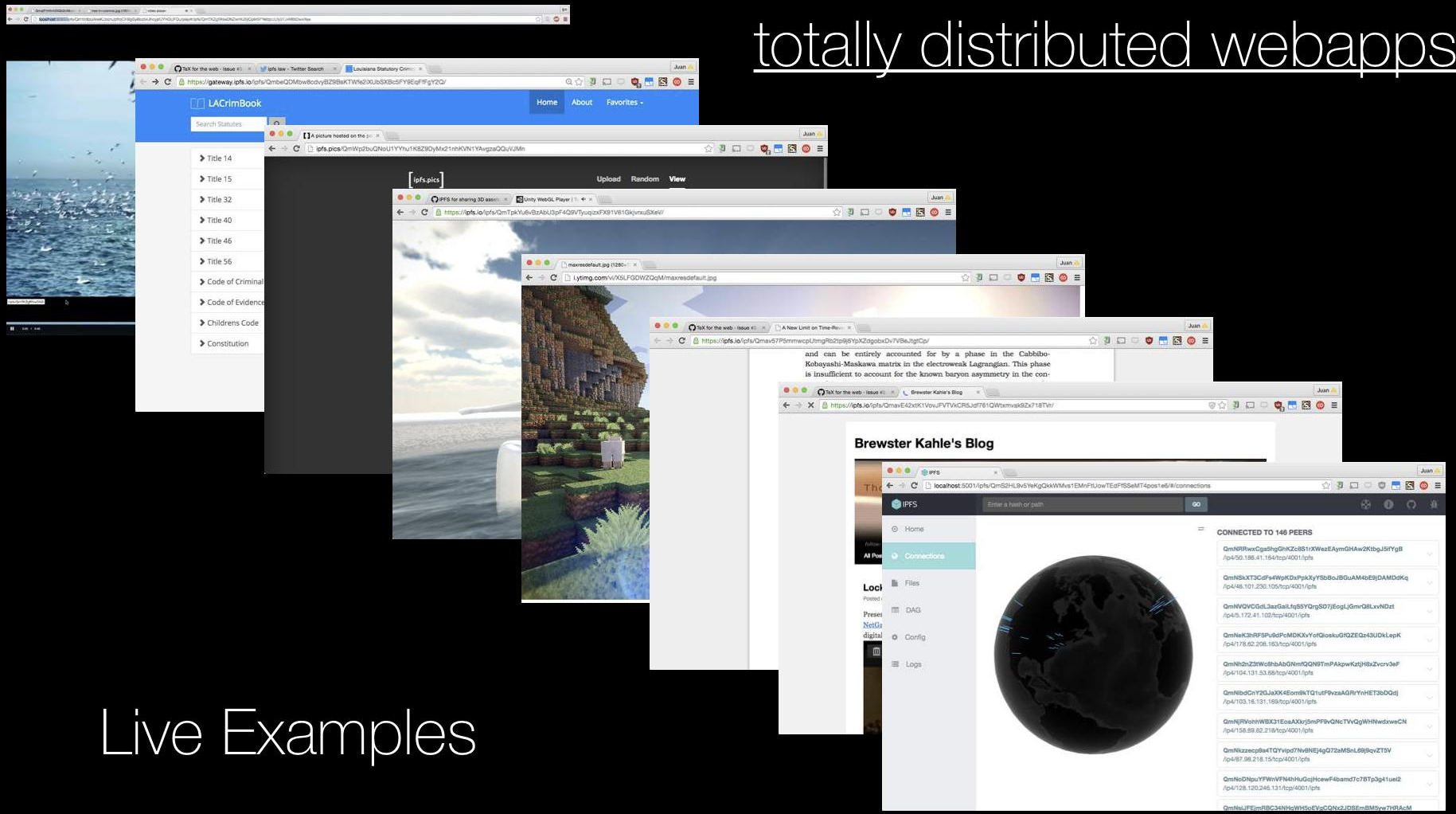
Live Examples

blogs and websites

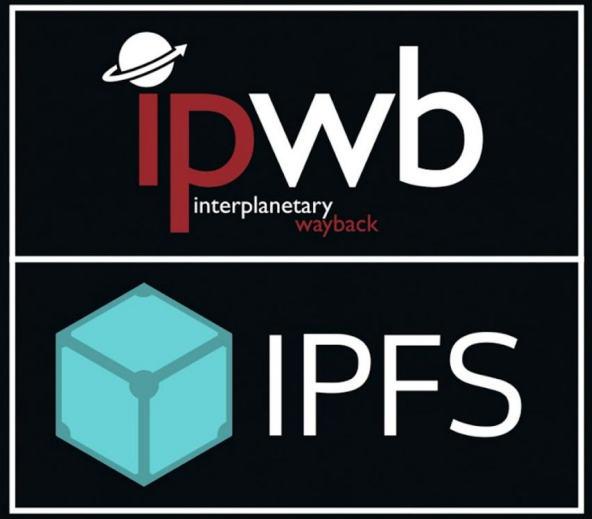


Live Examples

totally distributed webapps



Live Examples



- Distributed
- Offline
- Space savings
- Optimize bandwidth usage
- Improved resolution times
- and more..

ipwb

interplanetary wayback

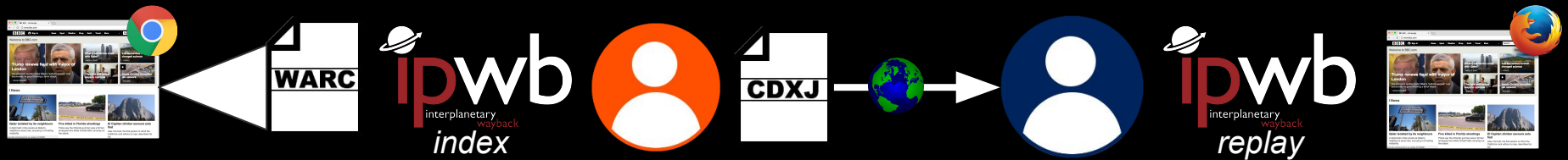
Motivation & Design

Motivation

- Persistence of archived Web data dependent on resilience of organization and availability of data
- Remove massive redundancy in Web archive files of exact duplicate content
- Determine feasibility of pushing WARCs into IPFS

Design

- Extending the CDXJ Format
- Indexing and IPFS Dissemination Procedure
- Replay and IPFS Pull Procedure



Design - CDXJ Format

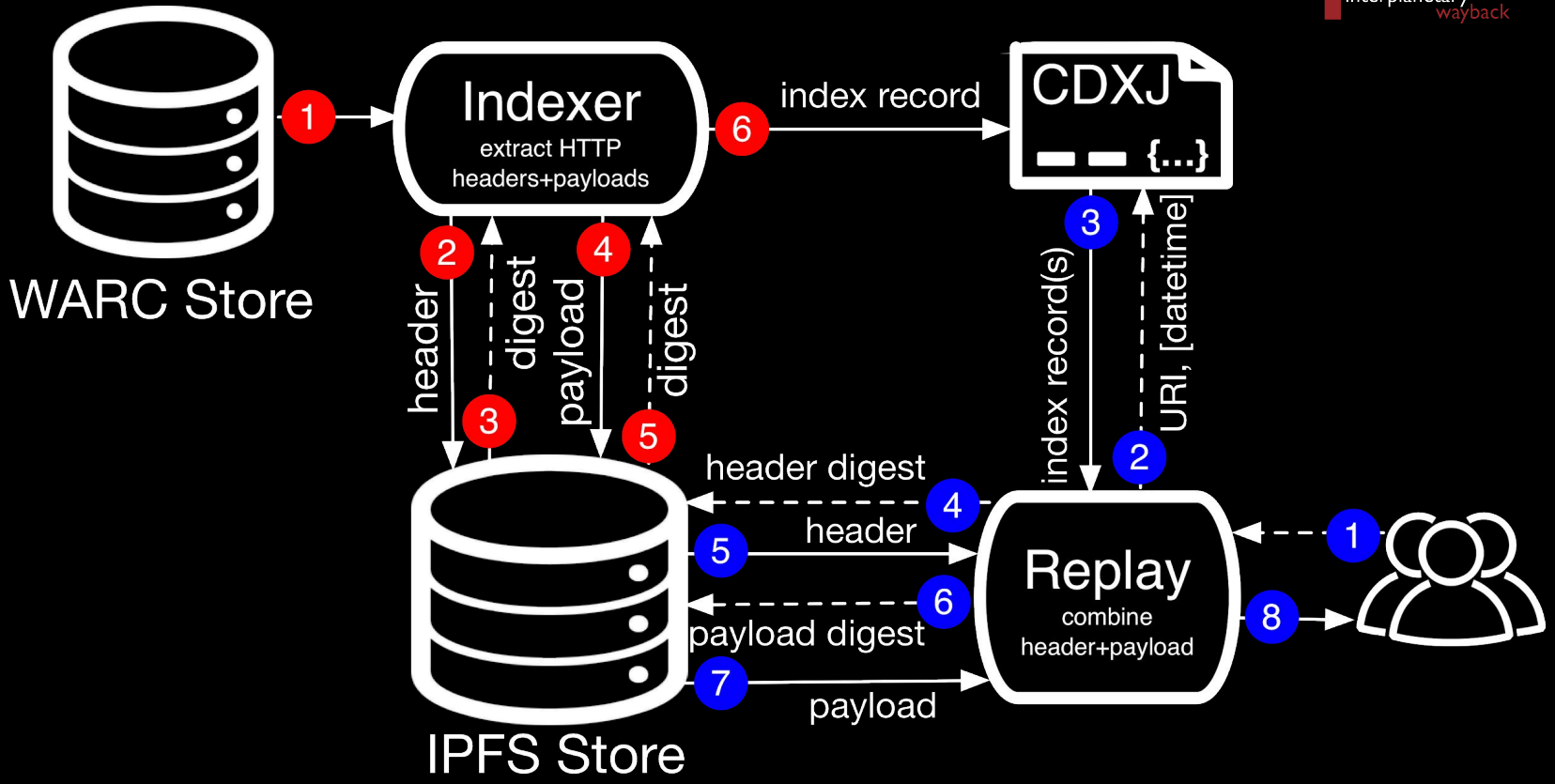
com,example)/index.html 20170301192639 {"mime_type": "text/html", "status_code": "200"}
com,example)/images/frog.png 20170301192639 {"mime_type": "image/png", "status_code": "200"}

See: https://github.com/oduwsdl/ORS/wiki/CDXJ
Alam et al. "Web Archive Profiling Through CDX Summarization", TPDL 2015

Design - CDXJ Format

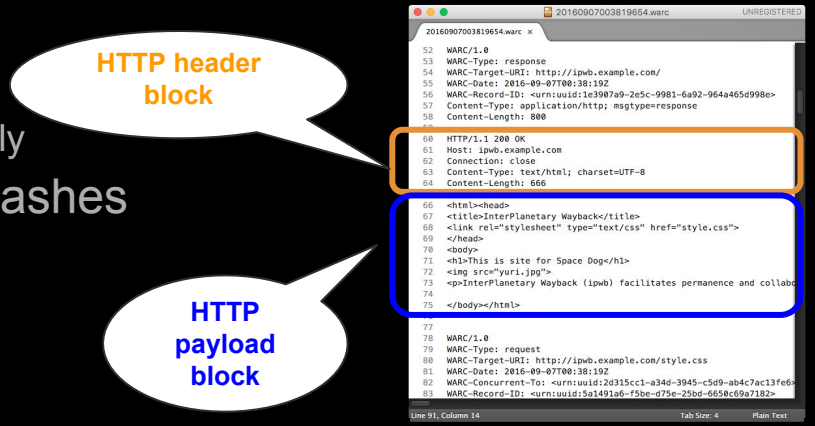
com,example)/index.html 20170301192639 {"locator": "urn:ipfs/QmPdyY6Pm66iWtGpTc7PqK11hvsnYSKMVL57G69RiNjGcm/QmNZ6mKSSAXAmXEocQj5gT4y4kdcr5D2C173ubWJ6PSKEZ", "mime_type": "text/html", "status_code": "200"}
com,example)/images/frog.png 20170301192639 {"locator": "urn:ipfs/QmUeko8zM7Xanwz6F9GtRH4rLAi4Poj3EMECGsci3BRQfs/QmPhMnX74cwqx2xgj9d3N3gTra8CzafXwSbUwU8xagMfqR", "mime_type": "image/png", "status_code": "200"}

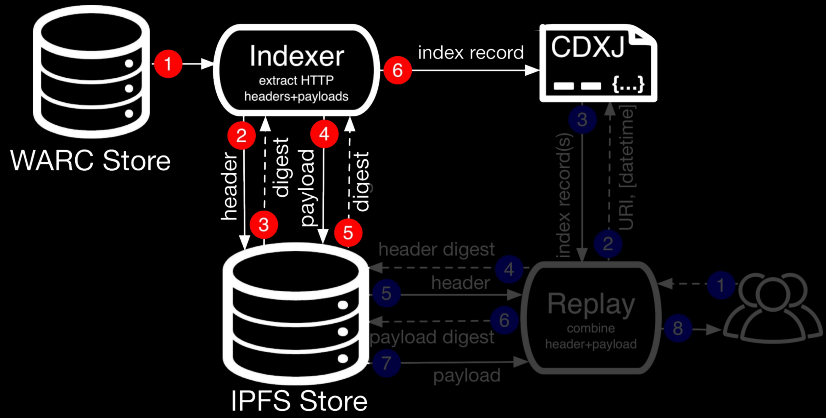
Design


WARC Store
1
Indexer
extract HTTP
headers+payloads
6
index record
CDXJ
{...}
2
header
3
digest
4
payload
5
digest
3
index record(s)
2
URI, [datetime]
IPFS Store
header digest
4
header
5
6
payload digest
7
payload
Replay
combine
header+payload
1
8

ipwb Design - “Indexing” Process

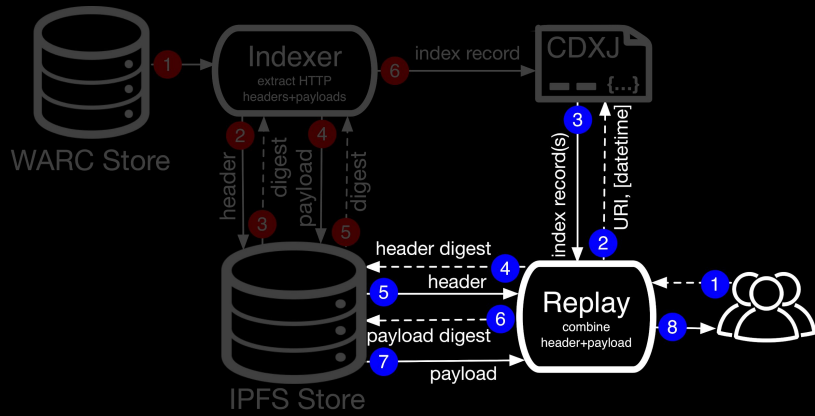
1. Extract HTTP Response from WARC
   - HTTP header and entity body (payload) separately
2. Push header and payload to IPFS, retain hashes
3. Construct CDXJ record containing:
   - URI of original resource (URI-R)
   - Datetime
   - Locator: urn:ipfs/headerHash/payloadHash
4. Repeat for each WARC-Response record
5. Save locally as CDXJ file



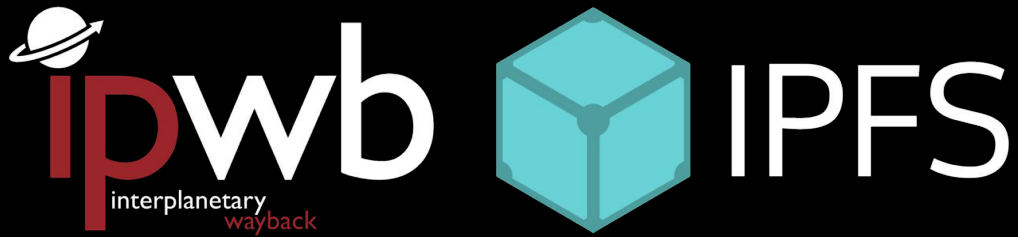


Design - Replay

1. Identify CDXJ line w/ URI-R + datetime
2. Fetch content for header and payload from IPFS using locator
3. Reassemble content into HTTP response, serve to browser
4. Repeat for each embedded resource requested



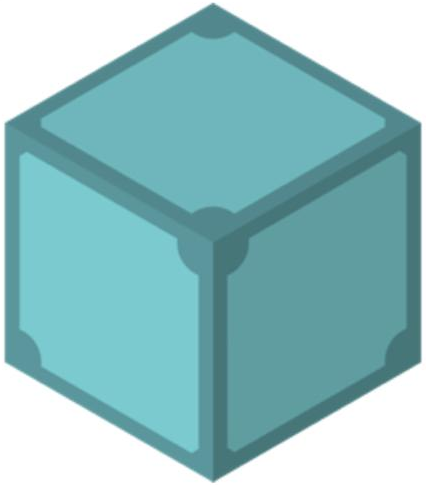
Advancements



IPFS

in the Browser

//github.com/ipfs/js-ipfs



IPFS .JS

The JavaScript implementation of the IPFS protocol.

made by Protocol Labs | project IPFS | freenode #ipfs

pm waffle | interface-ipfs-core API Docs | interface-ipfs-core Updates

build passing | PASSED | coverage 84%

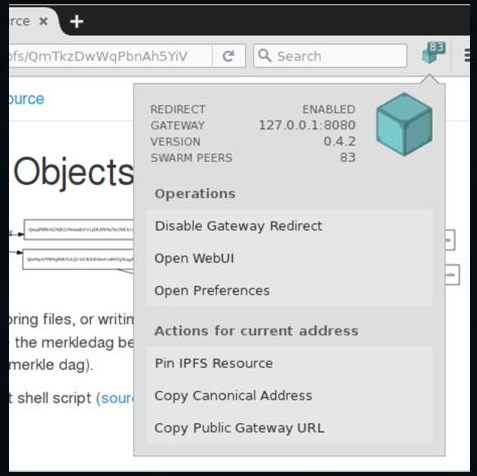
dependencies up to date | code style standard | standard-readme OK | npm >=3.0.0 | Node.js >=4.0.0

IPFS in the Browser

browser tab

```
// Create the IPFS node instance
const node = new IPFS()

node.on('ready', () => {
  // Your now is ready to use \o/

  // stopping a node
  node.stop(() => {
    // node is now 'offline'
  })
})
```

browser extension



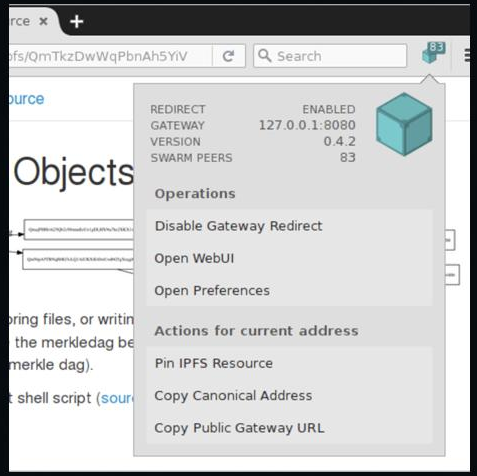
service worker

IPFS in the Browser

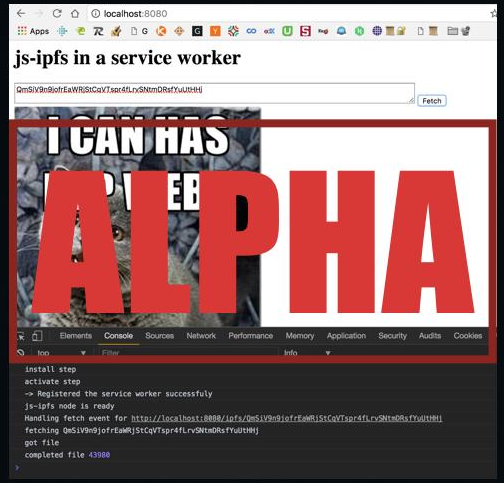
browser tab

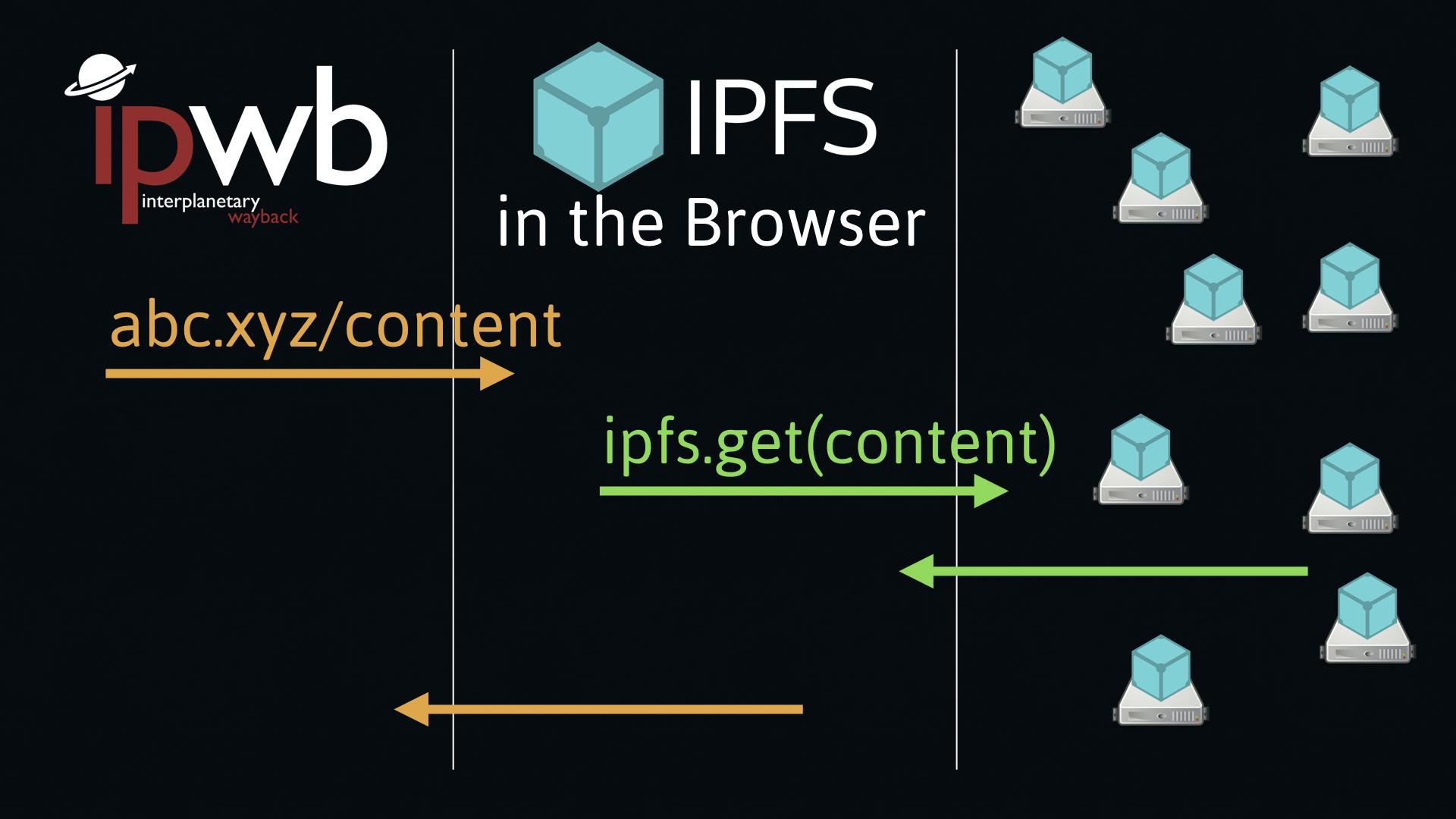
```
// Create the IPFS node instance
const node = new IPFS()

node.on('ready', () => {
  // Your now is ready to use \o/

  // stopping a node
  node.stop(() => {
    // node is now 'offline'
  })
})
```

browser extension



service worker




ipwb
interplanetary
wayback
IPFS
in the Browser
abc.xyz/content
ipfs.get(content)

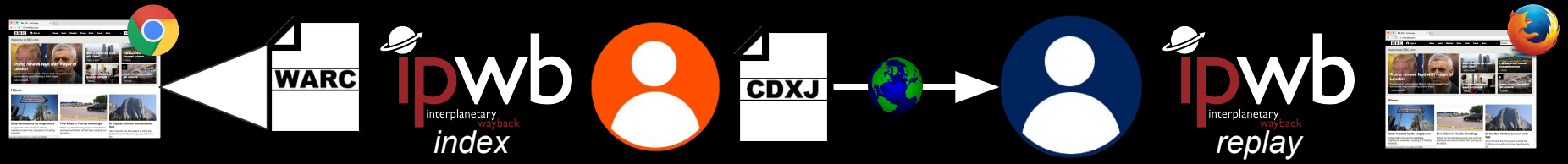
Privacy, Collaboration, and Security

- Encryption on indexing/dissemination, decryption on replay

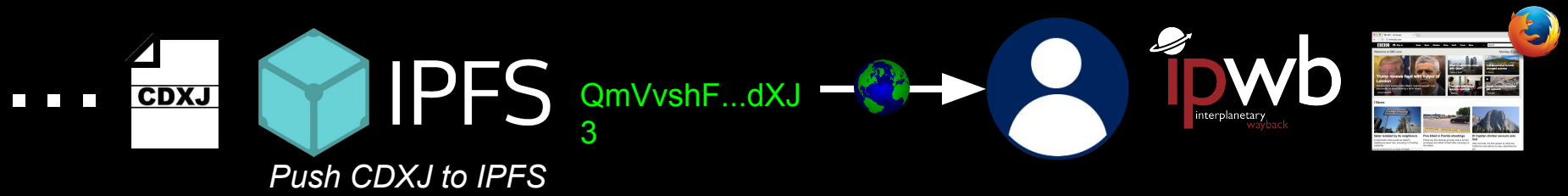
com,mywebsite)/photos/vacation 20170605083914 {
"locator": "urn:ipfs/QmdmV...P9Hf/QmRDB...1Bz2P",
"encryption_method": "xor", “encryption_key”:
“my#Gre4t#Encrypti0n#K3y!”, "mime_type": "text/html", "status_code": "200"}

Privacy, Collaboration, and Security

- IPWB CDXJs may be transferred for our users' replay



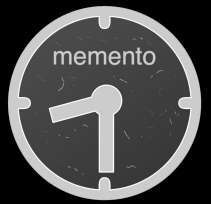
- CDXJ-by-hash recursive fetch/replay
  - Share hash of CDXJ then `$ ipwb replay hash` to replicate experience



Other ipwb Advancements

- Rerouting (instead of Rewriting) for Archival Replay*
  - IPWB replay registers ServiceWorker
    - Intercepts requests from archival replay to live Web
  - Prevents live Web from “leaking into” the archive on replay

- Memento Support
  - Replay system serves TimeMap, Timegate, and Datetime (memento) resolution endpoints
  - http://localhost/timemap/http://mywebsite.com/photos/vacation
  - http://localhost/memento/20170605092450/http://mywebsite.com/photos/vacation



* To be presented at JCDL 2017 in Toronto, Canada, June 19-23, 2017

A Collaborative, Secure, and Private InterPlanetary Wayback Web Archiving System Using IPFS

Mat Kelly
Old Dominion University
Norfolk, Virginia, USA
@machawk1

David Dias
Protocol Labs
Planet Earth
@daviddias

https://github.com/oduwsdl/ipwb

https://ipfs.io

w/ Sawood Alam, Michael L. Nelson, and Michele C. Weigle

IIPC Web Archiving Conference • June 15, 2017 • London, UK

Demo(s)