Content De-duplication for CDNi

http://www.ietf.org/id/draft-jin-cdni-content-deduplication-optimization-03.txt

WeiYi Jin (jin.weiyi@zte.com.cn)
Mian Li (li.mian@zte.com.cn)
Bhumip Khasnabish (bhumip.khasnabish@zteusa.com)

Wednesday, November 7, 2012
CDNi WG, IETF85, Atlanta

Outline

- Content de-duplication problem statement
- What is new in this version
- Proposed next steps for IETF CDNI WG
- Q&A and open Discussion

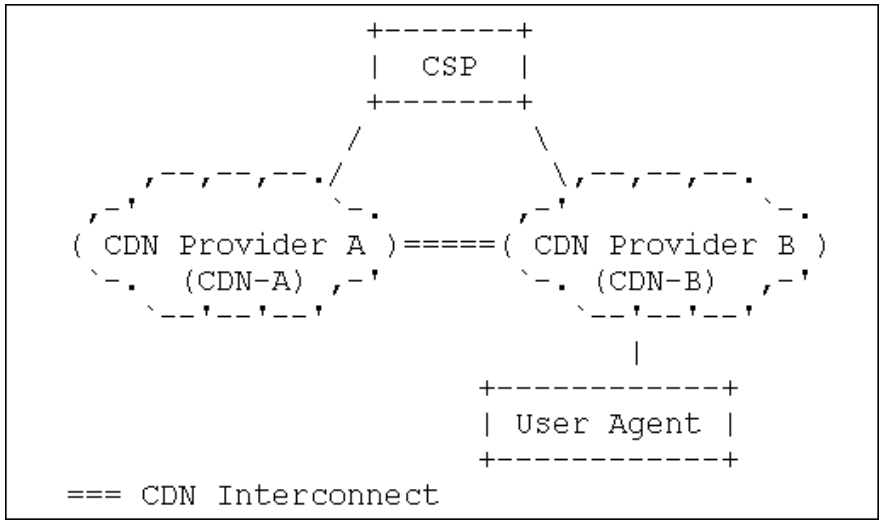
What's the Problem?

- Interconnected CDN with the same CSP;
- Interconnected CDN with the same CSP and with one dCDN;
- Cascaded CDN

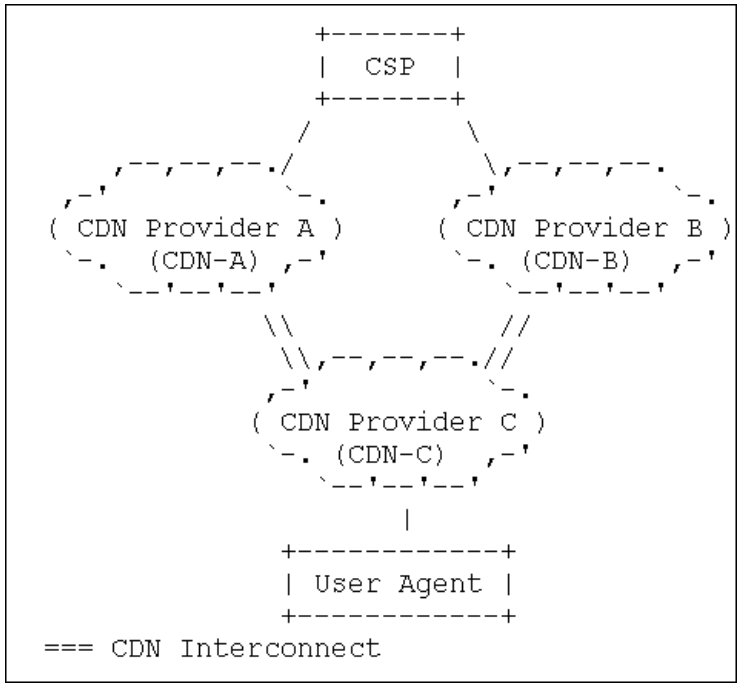
- Problem:
  - For a single given CSP, in some CDNi content via 'push' from one of the uCDN via pre-position procedure, and then may also request for downloading the same content from another uCDN via another pre-position procedure or upon user's content request.

- Reason for the Problem:
  - Currently the content is identified by a specific URL. Different CDNs may identify the same content with different URLs;
  - The URL used to identify the content may be changed during the redirection processes between CDNs;
  - No unique identification is used to identify a content item so far

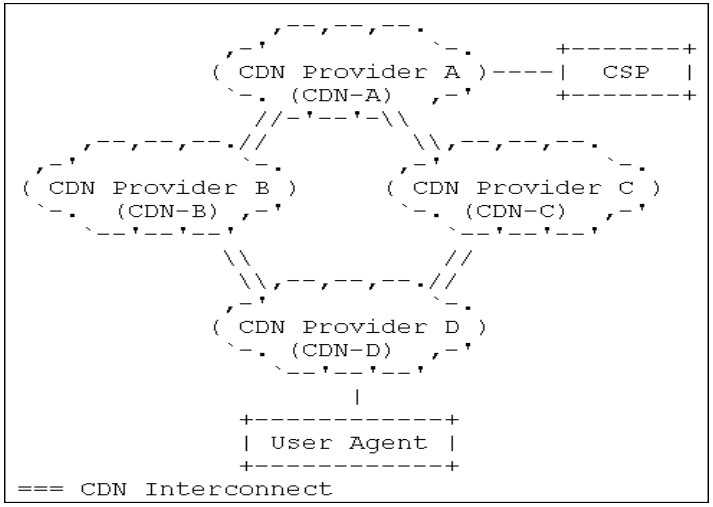
What's the Problem?



Scenario 1



Scenario 2



Scenario 3

What is New in This Version?

- Content duplication impact on current systems:
  - Survey exposes that up to 60% of the data stored in the application system is redundant and the proportion continues to grow along with the time moving forward;
  - For CDN, the duplication of the content delivered will:
    - increase the complexity of CDN management;
    - demand larger CDN storage capacity which would be a waste of facility and investment;
    - lead to inaccurate statistics of hot content, which consequently results in the inaccuracy of the arithmetic for the hot content dispatching in the CDN;
    - increase the latency for the CDN content access

- Current content de-duplication technologies:
  - Current content de-duplication technologies:
    - At present, data de-duplication technologies are widely used in the storage backup and archiving system;
    - Identical Data Detection: WFD, FSP, CDC, sliding block;
    - Resembling Data Detection: shingle, bloom filter, mode match
  - However,
    - Above technologies are applied under the prerequisite that the content is completed downloaded and stored;
    - for CDN, comparison of the content after downloading is a waste of transmission traffic and computation used for comparison;
    - Content duplication issue in CDNi reasons.

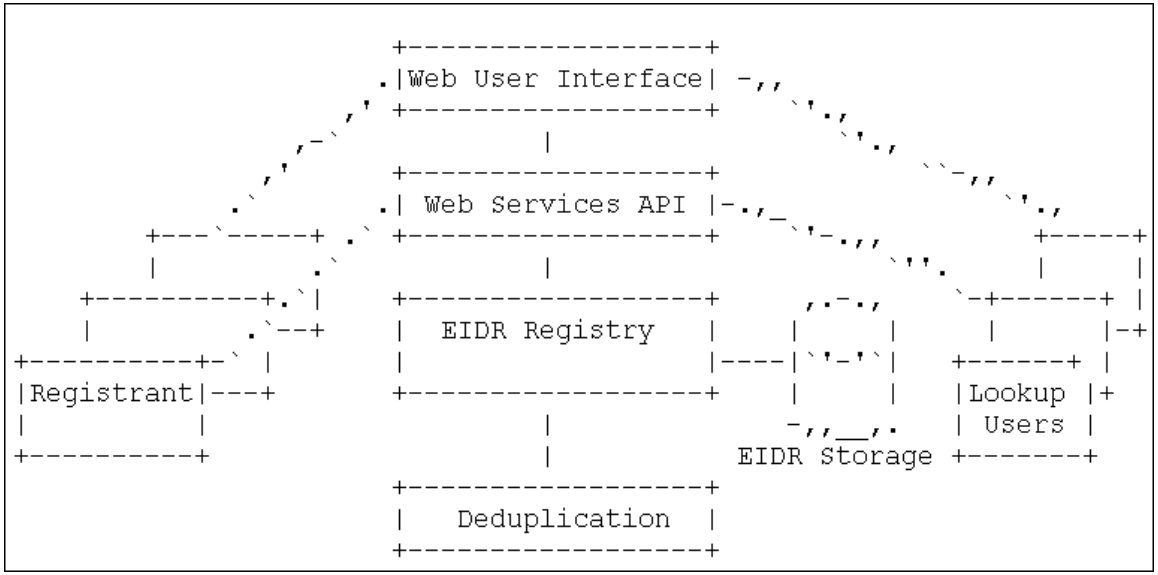
Identical Data Detection: WFD, FSP, CDC, sliding block;
Resembling Data Detection: shingle, bloom filter, mode match
However,
Above technologies are applied under the prerequisite that the content is completed downloaded and stored;

results from different

What is New in This Version?

- EIDR (Entertainment Identifier Registry), an industry non-profit organization, has already started the research and standardization of the globally unique content identifier and its generating mechanism

```
 Standard Prefix     Unique Suffix   check
for EIDR Registry   for each asset   digit
       |     | |                      |||
       10.5240/XXXX-XXXX-XXXX-XXXX-XXXX-C
```



What is New in This Version?

to content de duplication:

before it re-validates this content. The access to this content by other uCDNs will not be impacted;

- Purge --- If one of the uCDNs requests to purge certain content, this uCDN is not able to make any operation on this content, while the content itself is not deleted from the cache of dCDN. Only when all the uCDNs connected with this dCDN request to purge the same content and dCDN accepts all these requests will the content be purged from dCDN.

Proposed Next Step of CDNi WG

- Determine the necessity of content de-duplication in CDNi scope;
- Some technical discussion, maybe in mailing list;
- Further work or documents may include:
  - Include content de-duplication requirements in normative work;
  - Determine the CDNi content naming mechanism proposed in this version;
  - Enhance Metadata and Control models and interfaces to implement content de-duplication optimization;
  - Explore other possible solutions on content duplication issue

Q&A and Discussion