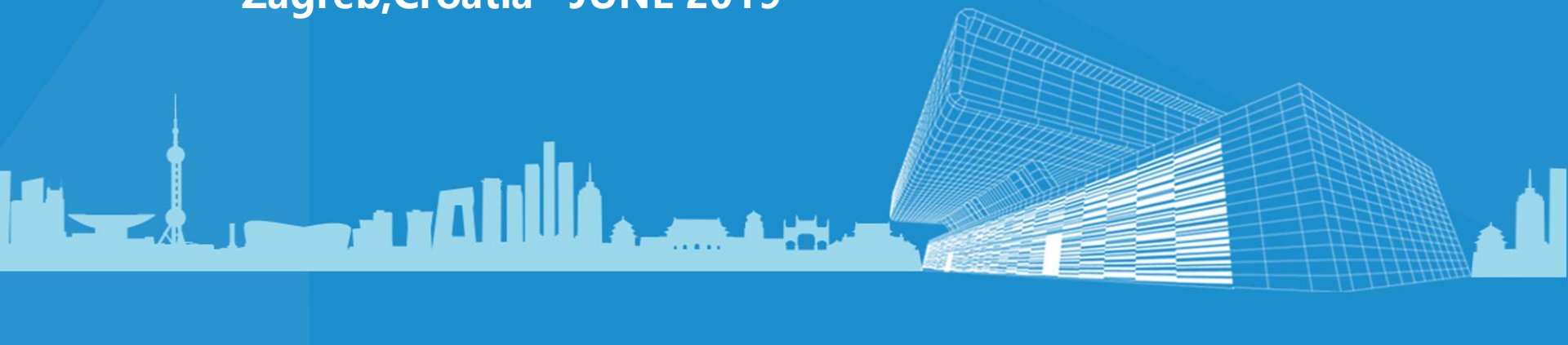
中国国家图书馆
NATIONAL LIBRARY OF CHINA

IIPC
INTERNATIONAL INTERNET PRESERVATION CONSORTIUM
netpreserve.org

The Key Technologies of Web Archiving Service Platform

JI Shiyan
National Library of China
Zagreb,Croatia JUNE 2019



Introduction

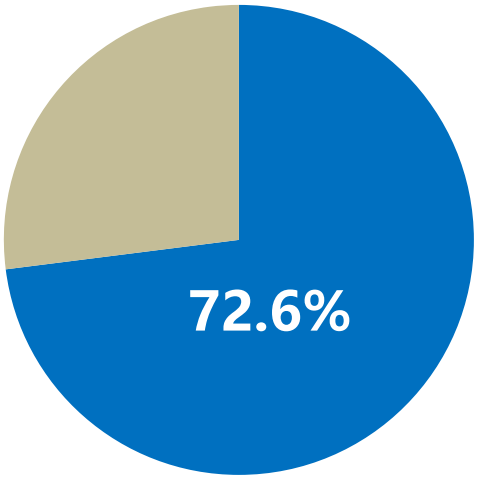
2003 Project Initiation

2007 A member of IIPC

2009 Web Archiving Center of National Library of China

2011 Nationwide project

2018 300 public libraries involved



Scope: government public information, key websites and webpages home and abroad

Web Archiving Service

Access at the NLC

Chinese Government Public Information Online

http://govinfo.nlc.cn

Web archiving on the basis of Heritrix

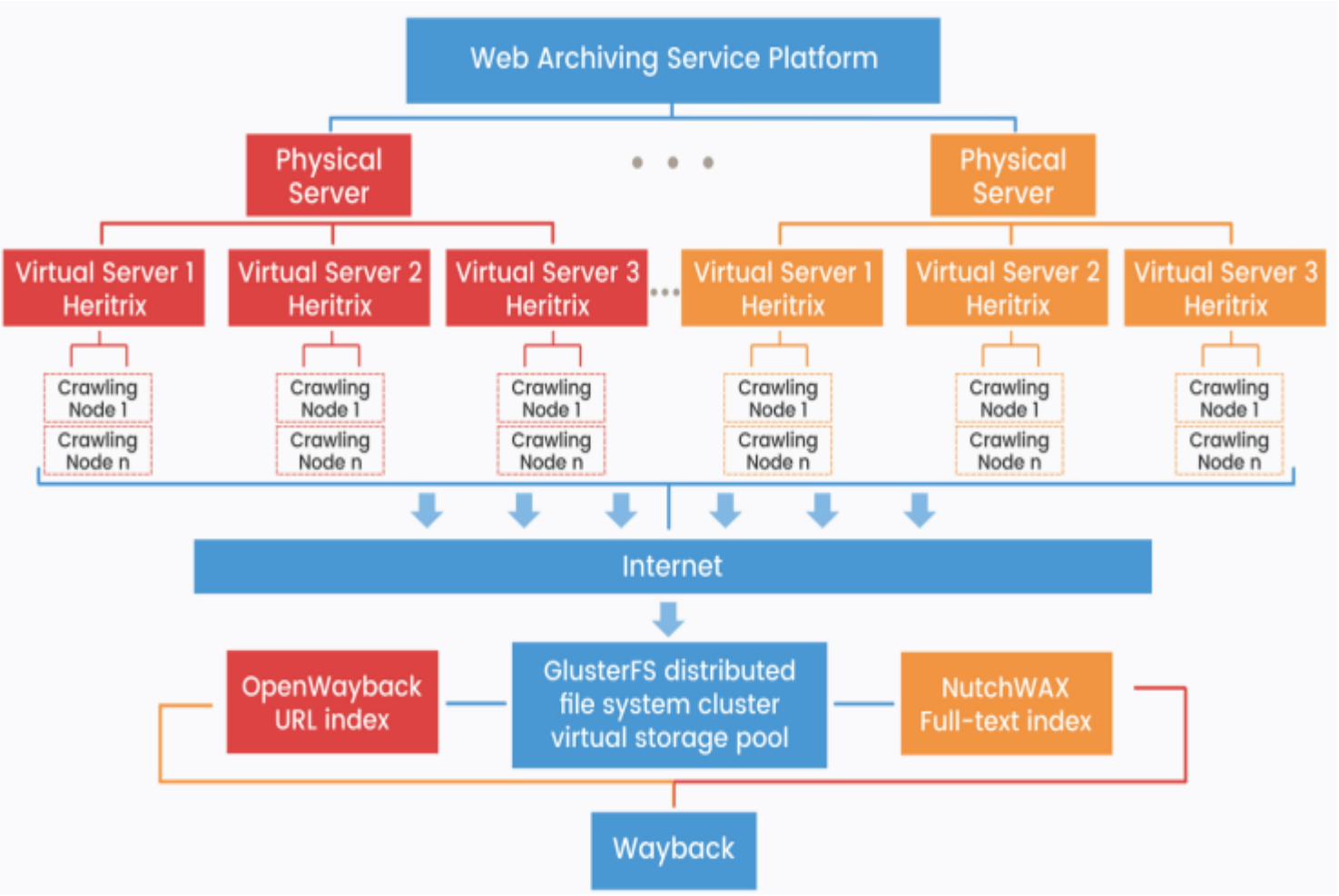
The NLC upgraded technology and developed the web archiving service platform in 2018

The explosive growth of web archives and the rapid development of network technology

More open, shared and compatible

Unite more libraries nationwide to participate in the project

Technology framework



A management terminal of B/S model

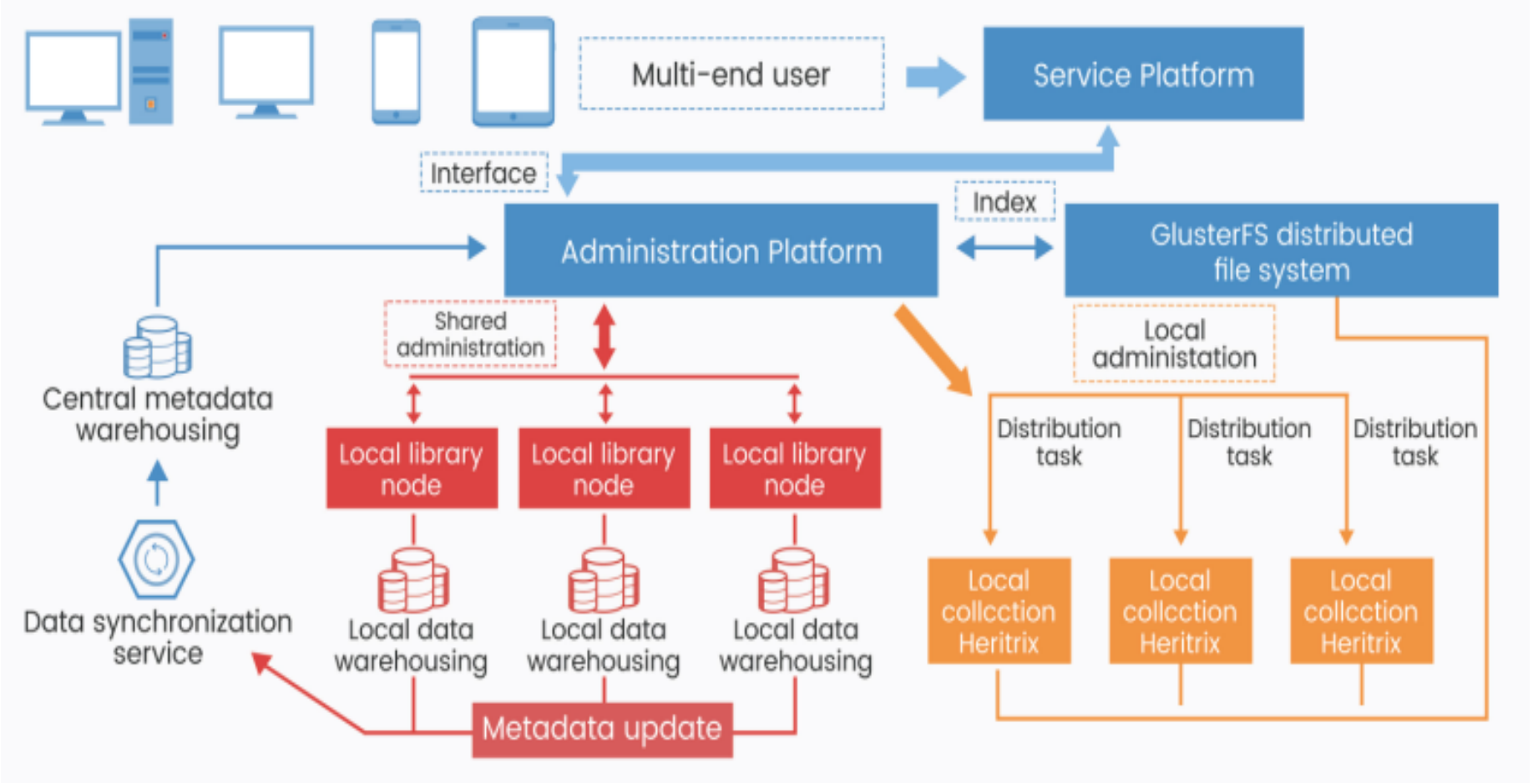
Virtualization technology

The DFS of GlusterFS

OpenWayback and NutchWAX components

Key Technologies

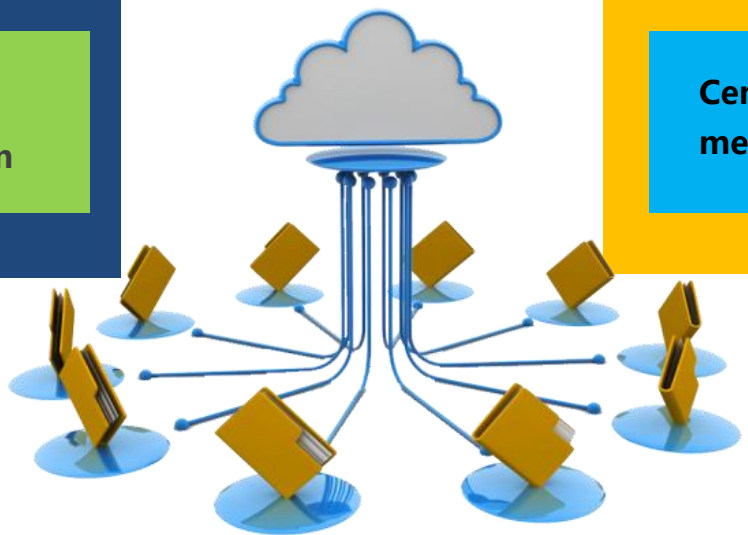
The Architecture of the Web Archiving Service platform



Key Technologies

Distributed acquisition and storage

- Distributed Storage Architecture
- Virtualization Technology
- Multithreading Heritrix
- GlusterFS
- MySQL
- Hashing Algorithm



Hierarchical Resource Sharing and Integrated Management

- The NLC owns a set of the framework of software& hardware
- Other node deploys the operational module as required
- Centralized management of metadata

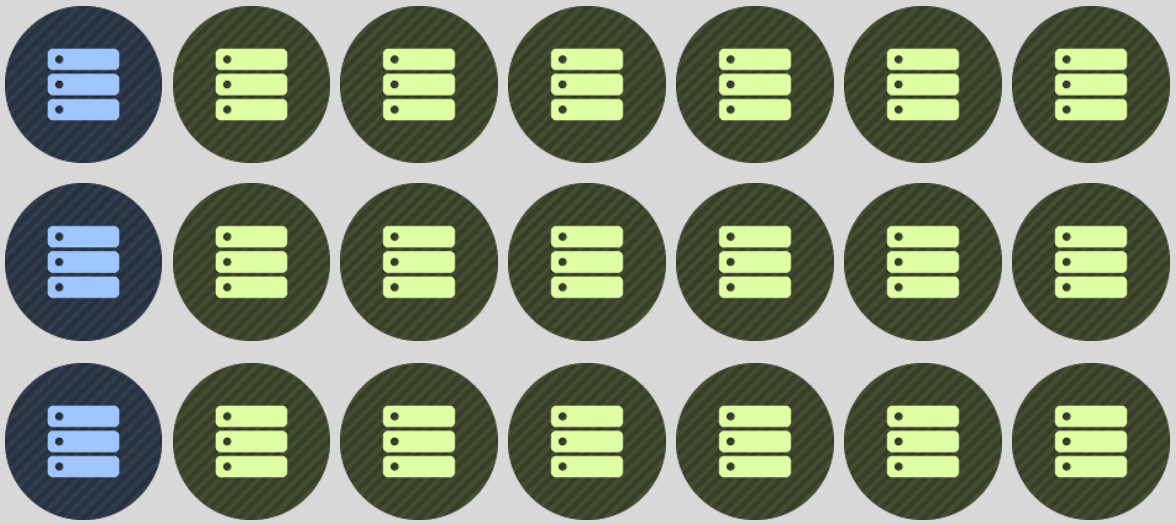
Conclusions

Home page of the administration platform

As of May 2019

- 7.8 TB of data
- 23720 URLs
- 403KB/s

Stable and Extensible Performance



211 crawling nodes

Web Archiving National Strategy Project

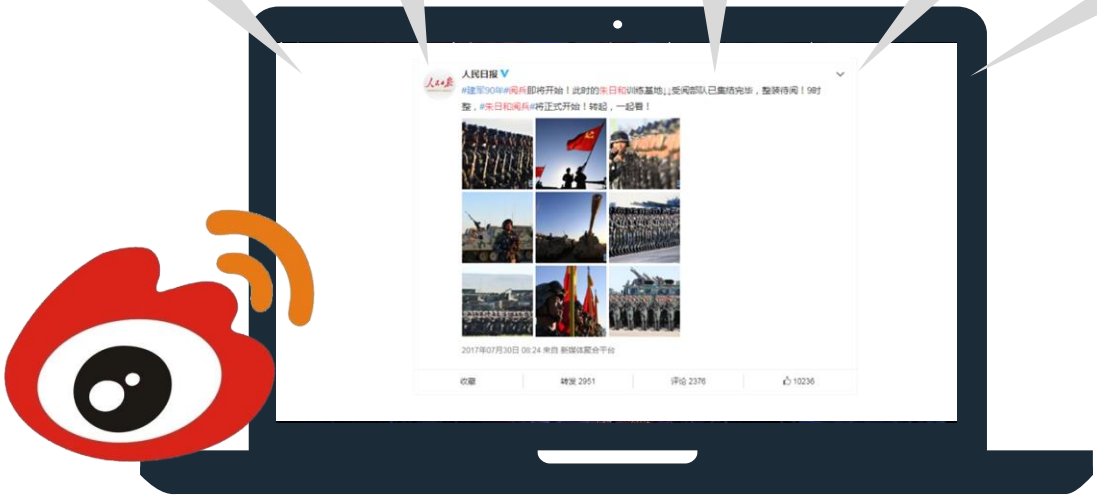
210 million news stories

200 billion public posts

50 billion pictures

400 million videos

500 billion comments



中国国家图书馆
NATIONAL LIBRARY OF CHINA

IIPC
netpreserve.org
INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

Thank You!

Contact Information:

JI Shiyan

Associate Research Librarian

Digital Resources Department of National Library of China

Email: jisy@nlc.cn