Tempest
IT makes sense

DEPOSIT
OF
DIGITAL PUBLICATIONS

ENVIRONMENTS USER INTERFACES BACKEND PROCESSES

www.tempest.technology

I. ENVIRONMENTS

Tempest

IT makes sense

ENVIRONMENTS: INTRODUCTION

MINIMAL HARDWARE CONFIGURATION

Tempest
IT makes sense

3x x64-based server with following characteristics

- 2 socket (12 cores per socket) CPU, 256 GB RAM
- 10 GbE LAN
- 2 x internal disk (RAID1) for virtualization hypervisor

All-purpose storage array for virtualization and processing of harvests

- 18 TB of usable capacity
- FC, iSCSI, NFS
- SSD cache
- RAID 1/5/6
- 10 000 IOPS

ENVIRONMENTS: INTRODUCTION

MINIMAL HARDWARE CONFIGURATION

Tempest
IT makes sense

NAS storage for harvested data

- 275 TB of usable capacity
- NFS support
- Active archive based on SATA disks

Virtualization platform

- Capable of running Linux guests (CentOS, RHEL, SUSE)
- Online migration of VMs
- Online storage migration of VMs
- High Availability for VMs in case of host failure

ENVIRONMENTS: INTRODUCTION

MINIMAL HARDWARE CONFIGURATION – BACKUP

Tempest
IT makes sense

LTO Library

- 1x LTO-7 capable tape library
- 1x LTO-7 FC drive

1x x64-based server (backup master server)

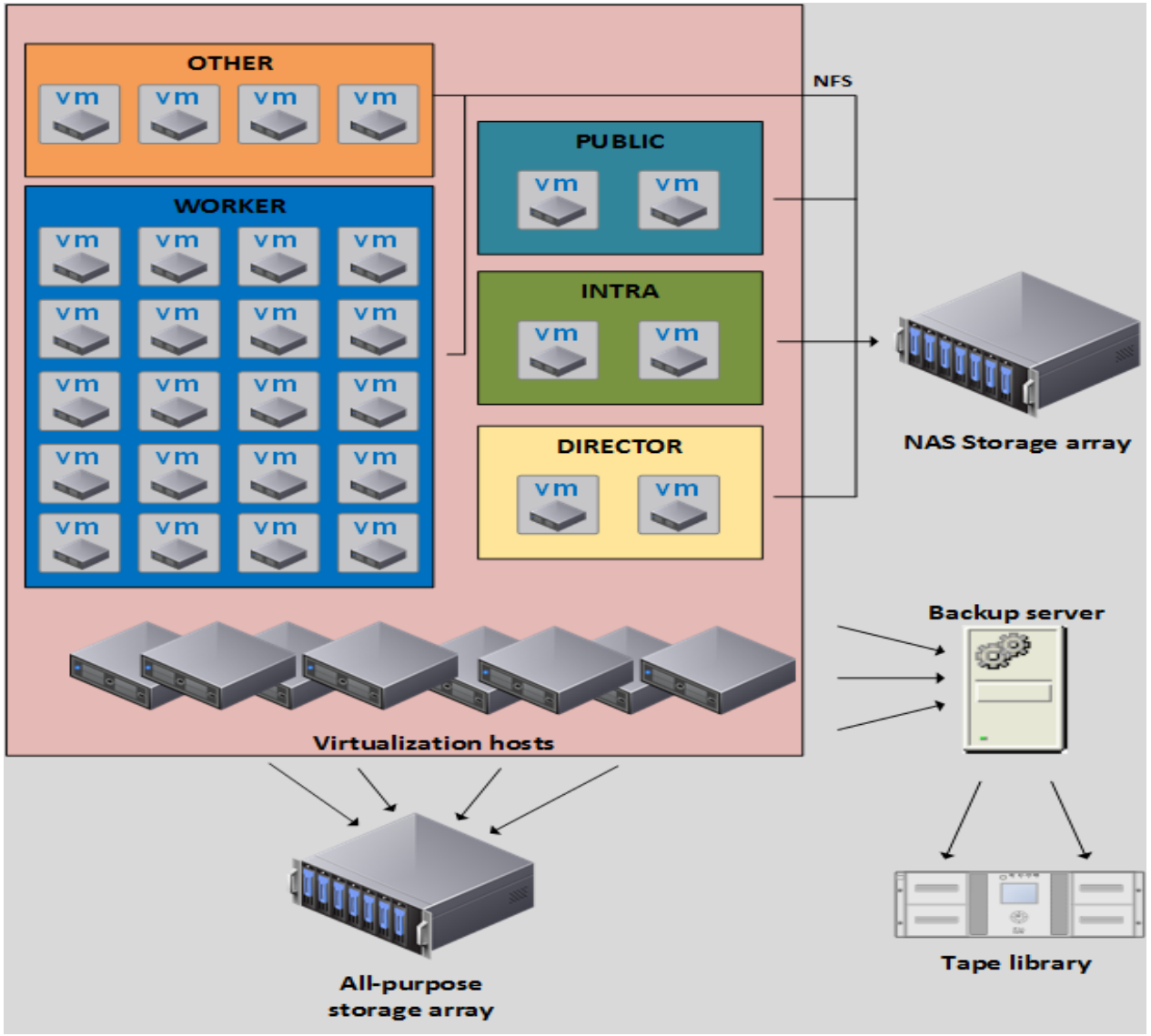
- 2 socket (10 cores per socket) CPU, 64 GB RAM
- 10 GbE LAN, 8/16 Gbit FC card
- 2 x internal disk (RAID1) for OS and backup software

Backup software with enterprise features

- Ability to backup virtualized environments (hypervisor aware)
- Snapshot management, agentless backup at the VM-host level
- Granular restores within VMs
- Support for Linux clients

ENVIRONMENTS: INTRODUCTION

SCHEMA OF MINIMAL INFRASTRUCTURE



Director server

- vCPU: 4, RAM: 16GB
- Director process coordinates background jobs

Worker server

- vCPU: 2, RAM: 8GB
- Worker process performs jobs according to messages from message queue

Public server

- vCPU: 2, RAM: 8GB
- Webapp-portal allows search for archived item

Intra server

- vCPU: 2, RAM: 8GB
- Webapp-intra allows plan harvests and administrate subjects

ENVIRONMENTS: INTRODUCTION

INTEGRATION TO EXTERNAL SERVICES

Tempest
IT makes sense

Invenio

- Service provider: University Library in Bratislava
- Obtaining metadata about items using ISSN, ISBN

SK Domain list

- Service provider: SK-NIC
- Obtaining list of SK domains

CDA

- Service provider: University Library in Bratislava
- Archivation of digital assets

ENVIRONMENTS: INTRODUCTION

3-RD PARTY COMPONENTS

Tempest
IT makes sense

Wordpress

- Open-source CMS under GNU GPLv2+ license.
- Provides general information about DDP (www.webdepozit.sk)

Heritrix

- Internet Archive's web crawler project under Apache license 2.0.
- Is managed by worker service

Openwayback (OWB)

- International Internet Preservation Consortium software for ‘play back’ archived websites.

ENVIRONMENTS: INTRODUCTION

3-RD PARTY COMPONENTS

Tempest
IT makes sense

SOLR

- Search platform under Apache license 2.0.
- Provides capability for :
    - search by some MARC metadata
    - search by SOLR expressions
    - full text search

RabbitMQ

- Message broker software under Mozilla Public License
- Provides temporary storage for job messages

PostgreSQL, OpenDJ, Spring Cloud

Relational database, LDAP service, Service discovery and configuration provider

ENVIRONMENTS: OVERVIEW

LIST OF ENVIRONMENTS

Tempest
IT makes sense

Production

- Datacenter of the ULB

Test 1 / Test 2

- Datacenter of the ULB

Docker, LXC

- Docker server in Tempest
- Computers of developers

Server \ Env.	Prod.	Test 1/2	Docker	LXC
Director	2	2	1	1+
Worker	239	10	1	1+
Intra	2	2	1	1+
Portal	2	2	1	1+

II. BACKEND PROCESSES

Tempest
IT makes sense

BACKEND PROCESSES

MAIN PROCESSES

Tempest
IT makes sense

Discovery

- Periodically collects metadata about domains
- Stores collected metadata as MARC fields
- Updates availability status of domains

Harvest

- Reads harvest jobs from RabbitMQ
- Creates instances of Heritrix for processing of harvests
- Manages Heritrix instances via API

OWB support processes

- Builds of OWB index files
- Generates OWB blacklists

BACKEND PROCESSES

SUPPORT PROCESSES

Tempest
IT makes sense

SK-NIC update

- Periodically downloads domains file from Slovak Top Level Registry
- Stores information about domains to a database

DROID signatures update

- Updates signature files for Digital Record Object Identification

CDA archivation

- Creates Submission Information Packages (SIP)
- Creates submission orders for SIPs
- Stores result of SIP processing to the database

III. HARVEST IN PRACTICE

Tempest
IT makes sense

HARVEST IN PRACTICE

HARVEST CONFIGURATION, WARC PRESENTATION

Tempest
IT makes sense

Harvest configuration >>

- Creation of harvest template
- Harvest scheduling

SOLR search >>

- Search separately for WWW and E-Born
- Instant search by Conspectus, Facets, Fulltext search
- Advanced search by MARC field, SOLR expression

OWB presentation

- Background blacklist generation
- Content access by access type (Public, Library, Curator)

IV. PRODUCTION ISSUES

Tempest
IT makes sense

PRODUCTION ISSUES

Tempest
IT makes sense

Quality of data

- Too complex crawler configuration
- Cyclic links in WARC file
- Deduplication strategy

Harvest issues

- Complaints from content provider
- Risky harvest without obeying robots.txt
- Ban from Web housing providers for DOS attack

Telefón
+421 (2) 502 67 111

Fax
+421 (2) 502 67 100

Informácie
info@tempest.sk

Obchod
obchod@tempest.sk

www.tempest.sk

TEMPEST a. s.
Galvaniho 17 / B
821 04 Bratislava 2
Slovenská Republika

THANK YOU FOR YOUR ATTENTION

Tempest
IT makes sense