cesnet

SLUŽBY DATOVÝCH ÚLOŽIŠŤ CESNET

Michal Strnad

CESNET

31.5. 2023

Prezentace

- Přehled služeb pro ukládání dat: ownCloud, Filesender, objektové uložiště – S3 a RBD
- Praktické příklady použití S3 pro spolupráci vědeckých týmů pro práci s velkými a citlivými/medicínskými daty
- Praktický příklad architektury pro zálohování ústavu AV ČR do CESNETu

IRQ možno během prezentace, ideálně ale až na konci

- Přehled služeb pro ukládání dat: ownCloud, Filesender, objektové uložiště – S3 a RBD

Datová úložiště

- v současnosti 6 datacenter s celkovou hrubou kapacitou 129 PB
  - většina kapacity dostupná na objektové technologii – S3, RBD
  - menší diskové pole (souborové úložiště) 9 PB – klasické služby NFS, ftp, samba
  - přechod na objektové technologie
- elementární případy užití
  - archivy, zálohy, sdílení dat, synchronizace dat
- velmi důležité je téma dat opatřených metadaty
  - FAIR, Open Science a další moderní trendy
  - Národní repozitář

- sync'n'share („cloudové“) úložiště
  - desktopová aplikace (Windows, Linux, Mac OS X)
  - mobilní aplikace Android, Apple
  - a webové rozhraní
- data se synchronizují přes úložiště
  - na počítači jsou i lokálně, na mobilním zařízení při otevření
- data lze sdílet
  - konkrétní osobě nebo „kdo zná odkaz“
- registrace federací na `https://owncloud.cesnet.cz`
- standardní limit 100 GB na uživatele
- 20.000 registrovaných uživatelů

Jednorázové zaslání souboru

- FileSender: webová služba pro jednorázový přenos (velkých) souborů
- `https://filesender.cesnet.cz`
- alespoň jedna strana komunikace musí být oprávněný uživatel infrastruktury
  - autentizace federací eduID.cz
- oprávněný uživatel může nahrát soubor a poslat druhé straně oznámení
- lze poslat komukoli pozvánku
- lze uploadovat i pomocí příkazové řádky (python klient)

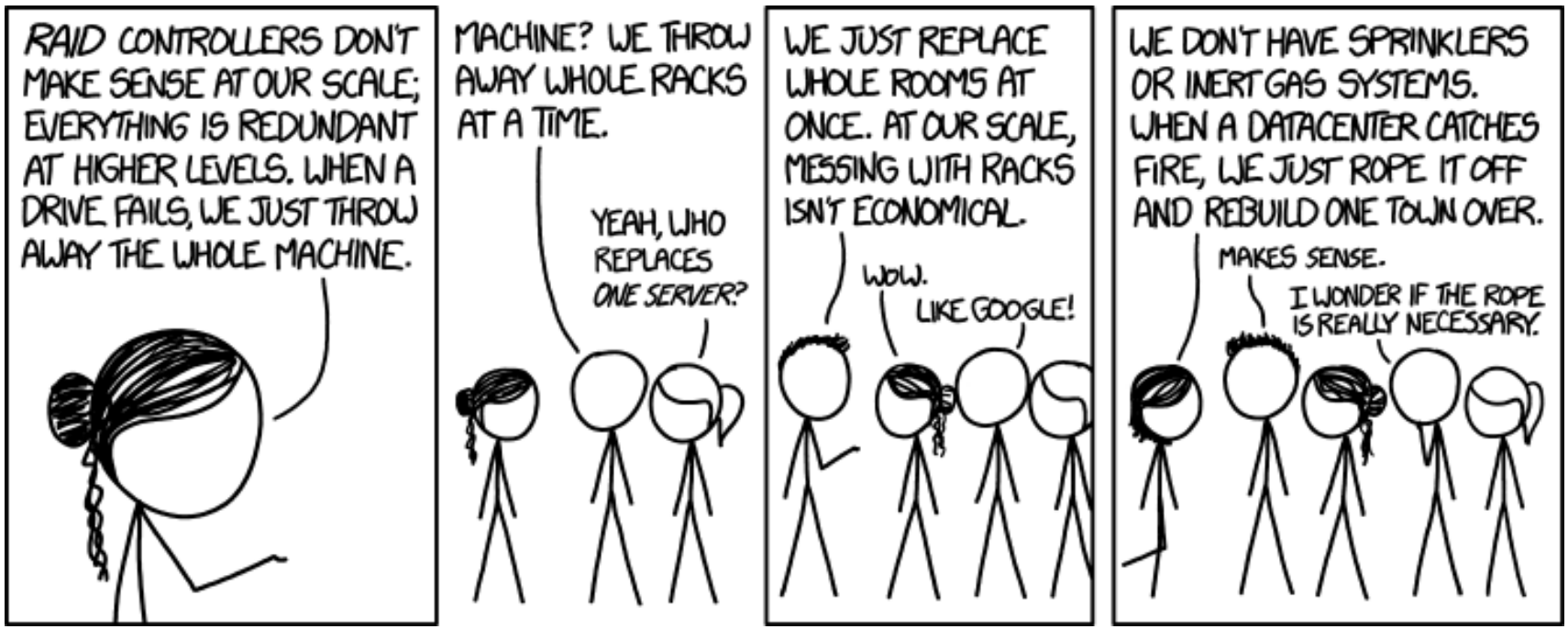
Objektové služby

- Objektová uložiště vs klasický souborový systém
  - Flat struktura
  - Rozšířená metadata pro vyhledávání, organizaci a správu obsahu
  - Rozšiřitelnost a škálovatelnost - dynamicky rozšiřování bez omezení hierarchie adresářů
  - Redundantní kopie (replikace)
  - Jemnější kontrolu přístupu k objektům a metadatům
  - Ceph, GlusterFS, IBM Spectrum Scale, Scality ...

Objektové služby

- Objektová uložiště - Ceph
  - Objektově orientované uložiště
  - Cluster sám udržuje minimální počet nastavených replik/EC
  - Větší odolnost proti výpadku serveru, racku, či dokonce celého datacentra
  - Protokoly S3/Swift, CephFS a RBD
  - OSD, MON, PG, CRUSH

Objektové služby



Objektové služby

- Objektová uložiště - Ceph
  - každý cluster má desítky serverů a tisíc až dva tisíce disků (MTFB)
  - server - 4x10 Gbps / 2x25 Gbps, 300 - 750TB, min. 256 GB RAM
  - cluster připojen N x 100 Gbps

Objektové služby - S3

- Služba S3
  - připojení pomocí https, GUI nebo command line klienti pro Windows i Linux
  - v porovnání s HSM systémy řádově rychlejší, zejména pro čtení
  - jednoduché sdílení pro autentizované i neautentizované uživatele
  - dostupné odkudkoliv z internetu
  - umožňuje široké spektrum nasazení pro velmi specifické případy užití
  - využitelné pro různé typy uživatelských aplikací pro distribuci dat
    - data se neukládají na server s aplikací, ale přímo na objektové úložiště
- S3 je za nás preferovaný protokol

Objektové služby - S3

- Služba S3 - terminologie
  - tenant
  - access a secret klíče
  - bucket
  - endpoint

Objektové služby - S3

- Služba S3 - terminologie
  - pre-signed objekt
  - jedinečné časově omezené URL na základě metadat objektu a přístupových klíčů
  - objekt již následně nelze přepsat
  - po vygenerování presigned URL adresy, nemůžete měnit její životnost
  - pozor na Google, DuckDuckGo, ... Shodan

Objektové služby - S3

- Služba S3
  - s3cmd, aws-cli, s5cmd

Reklamní vložka na AWS CLI plugin

- `github.com/CESNET/aws-plugin-bucket-policy`
- je potřeba to řádně otestovat

Národní datový repozitář

- repozitář: úložiště pro data opatřená metadaty
  - Findable, Accessible, Interoperable, Reusable
- aktuální stav – pilotní provoz
  - prototyp repozitáře pro data
  - záznamy určené k publikaci kontrolujeme a publikujeme ručně (NTK)
  - https://data.narodni-repozitar.cz/
  - lze vytvořit komunitu uživatelů
  - tj. skupinu se jmenovanými správci záznamů
  - lze přidělit persistentní identifikátory (DOI)
  - nutností je alespoň základní metadatový popis záznamů
  - uživatel řídí rozsah zveřejnění záznamu
- informace včetně žádosti o přístup naleznete na:
  https://du.cesnet.cz/cs/navody/narodni_repozitar/start

Repozitářové platformy

- plány podpory FAIR dat 2022–2029:
  European Open Science Cloud (EOSC)
- Národní repozitářová platforma
  - systém pro vytváření instancí/tenantů repozitářů na míru
  - pro potřeby vědeckých komunit, institucí, …
  - navázaný na další služby e-infrastruktury
- „pomocné systémy“ repozitářové platformy
  - Národní metadatový adresář
    - centrální vyhledávací bod, agregátor metadat
  - správa přístupů, řízení uživatelů
  - spolehlivé ukládání dat (geografické repliky, kontroly)
  - nástroje pro přenosy dat
  - obecná úložiště pro nestrukturovaná data

Shrnutí

- preferované služby S3 a RBD v různých kombinacích zcela dle požadavků uživatele
- služby jsou modulární a dají se spojovat k vytvoření workflow přesně na míru
- Národní repozitář
  - pilotní provoz datového repozitáře na vyzkoušení
  - žádost dostupná na: **https://du.cesnet.cz/cs/navody/narodni_repozitar/start**
- Podpora FAIR dat 2022-2029
  - Národní repozitářová platforma
  - Národní metadatový adresář
  - Podpůrné služby – přenosy dat, kontrola integrity, řízení přístupů

Otázky?

- Praktické příklady použití S3 pro spolupráci vědeckých týmů pro práci s velkými daty (nejen citlivými/medicínskými)

Objektové služby - S3

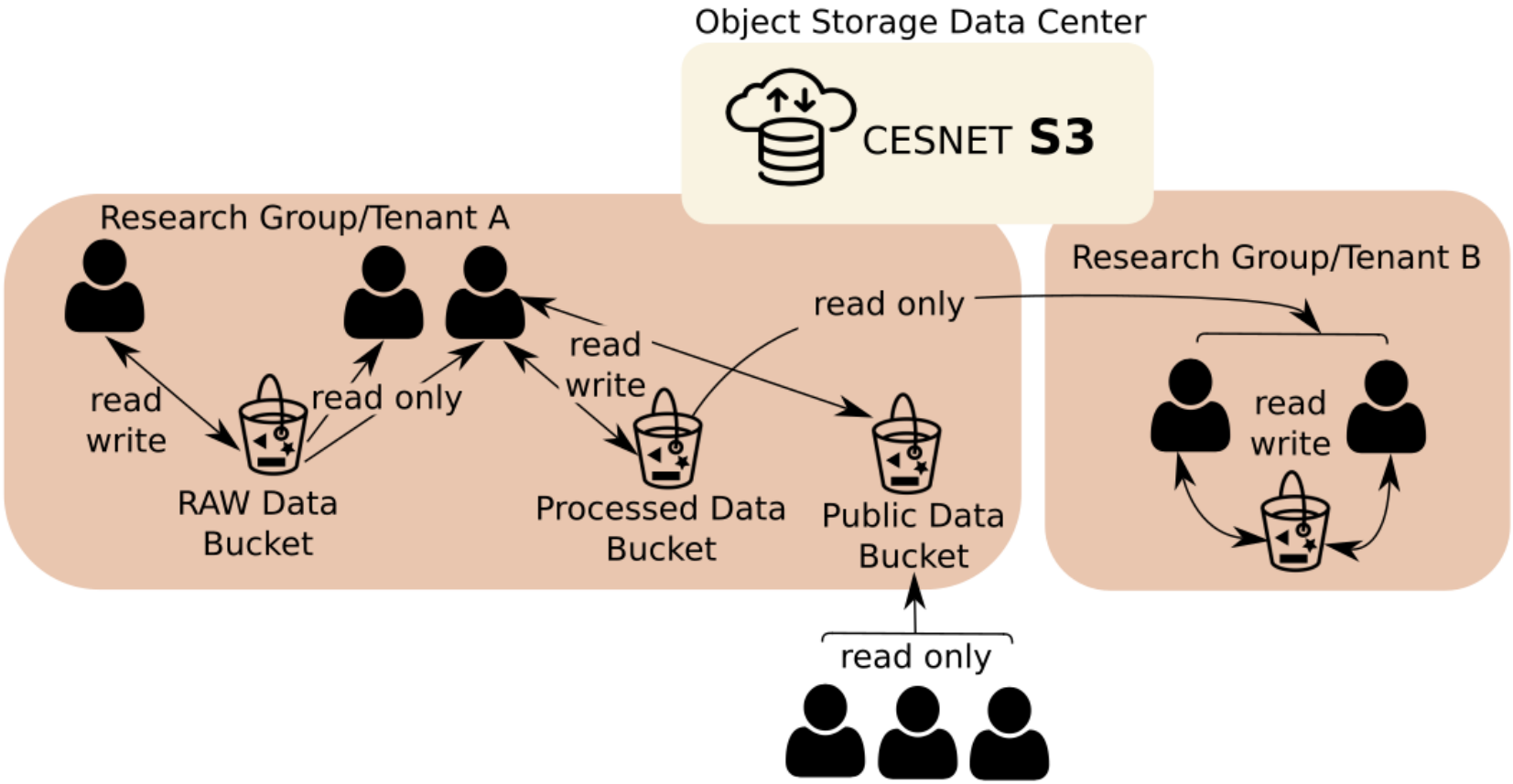
- Anonymizovaná, citlivá data ...
  - neposkytujeme server-side šifrování
  - šifrování je nutné provádět na klientské straně
  - WORM model

Objektové služby - S3

- write-once-read-many (WORM) model
  - Prerekvizita - verzování na bucketu (při vytvoření)
  - object lock = immutable (legal hold, retention period)
  - governance vs compliance mode

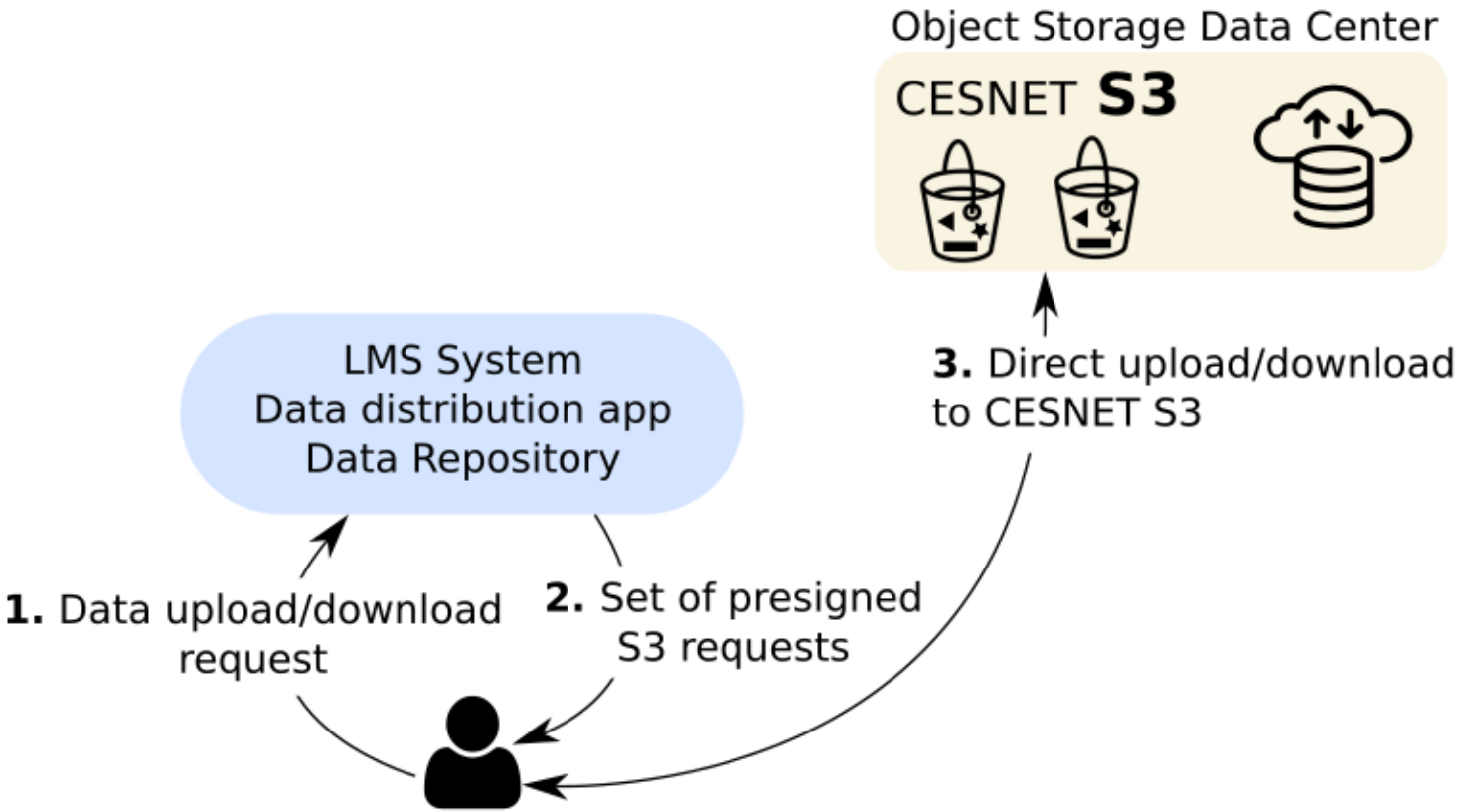
Popis případů užití S3 #1

Projektové workflow pro sdílení a distribuci dat

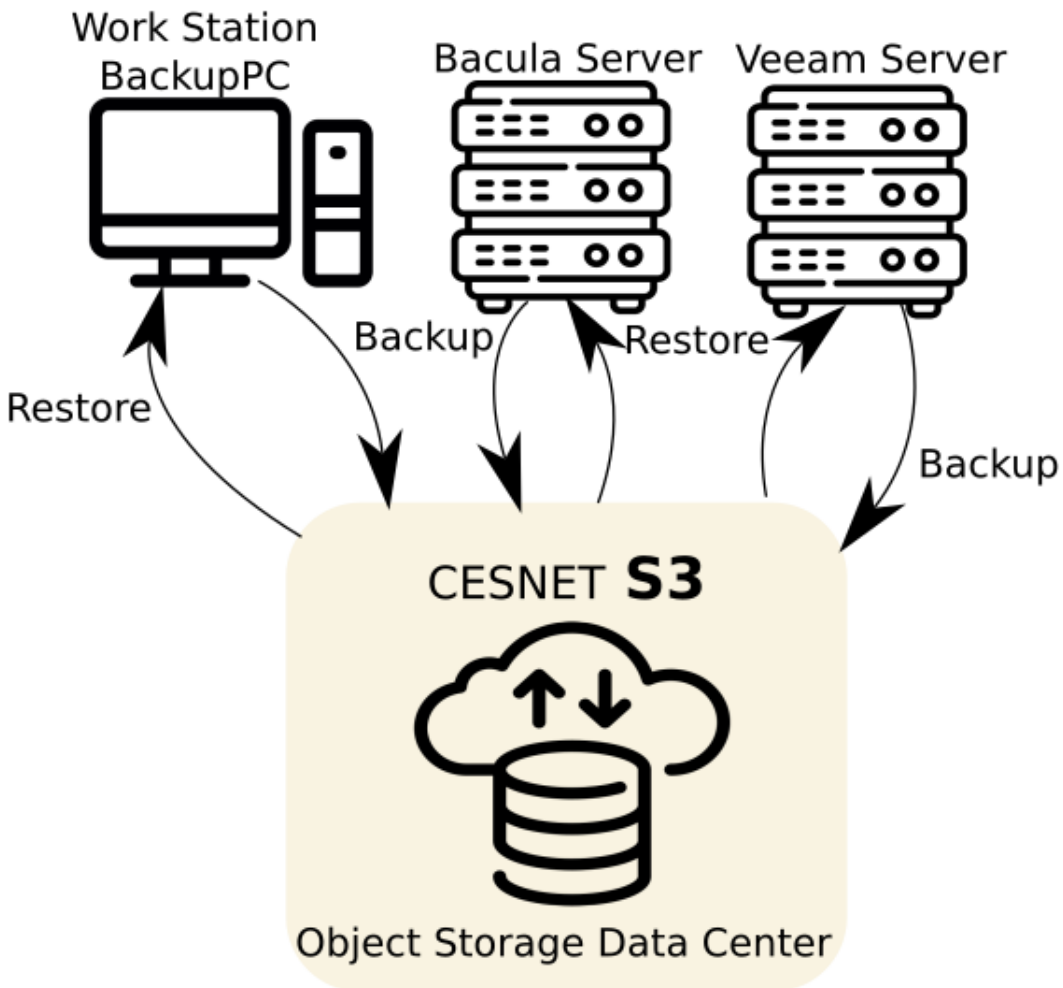


Popis případů užití S3 #2

Přímý upload/upload dat do aplikace (Learning Management Systems, Repozitáře...)



Popis případů užití S3 #3



Popis případů užití S3 #4

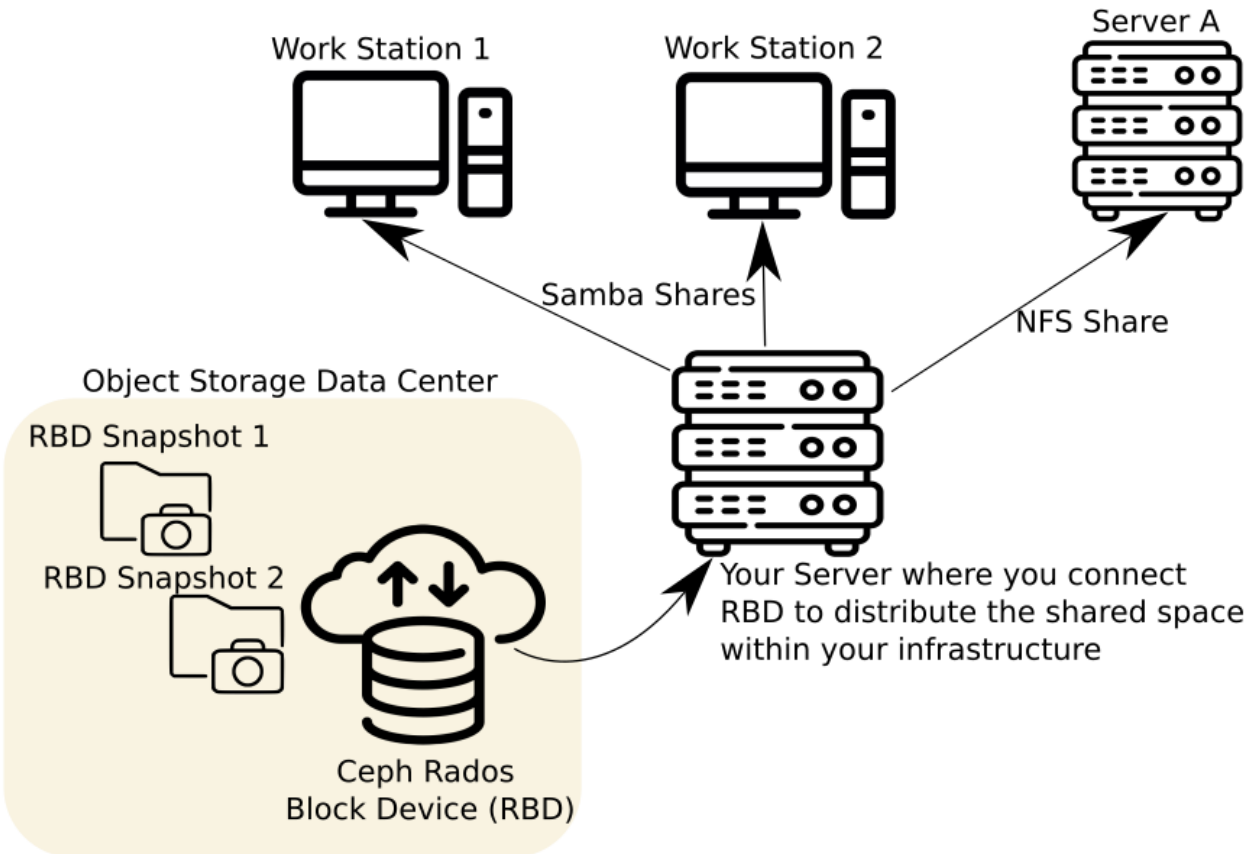
Stagging pro výpočty

Objektové služby - RBD

- Rados Block Device – blokové zařízení
- podobné jako síťový disk, diskové pole
- nutné připojit do infrastruktury pomocí Linuxového stroje
- připojení je možné z definovaného IP rozsahu – firewall
- nastavení zcela dle preferencí správce - file systém, client side šifrování...
- pozor na připojení přes více strojů zároveň
- pozor na saturaci sítě (VRF)

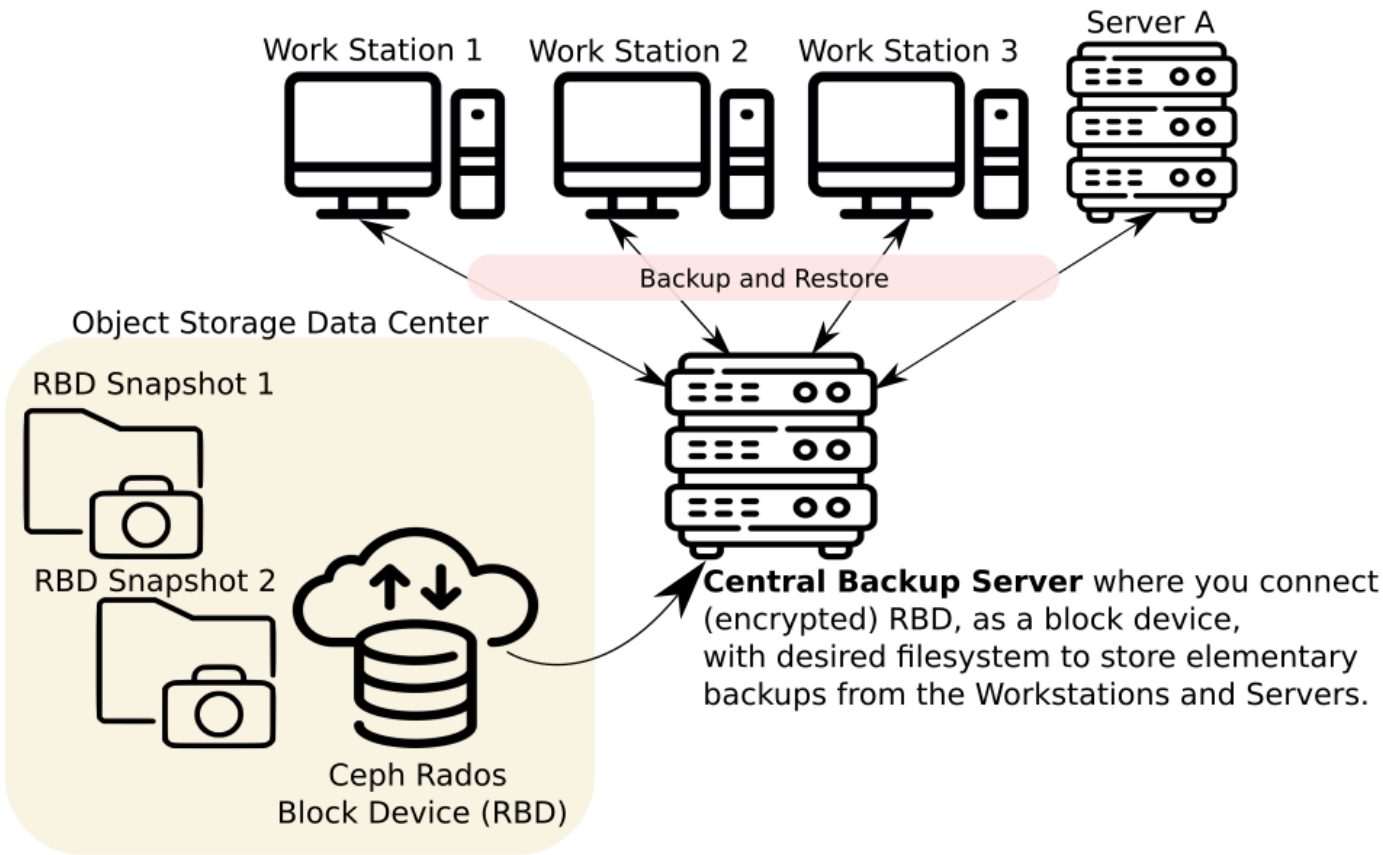
Popis případů užití RBD #1

Redistribuce úložiště dovnitř vlastní infrastruktury

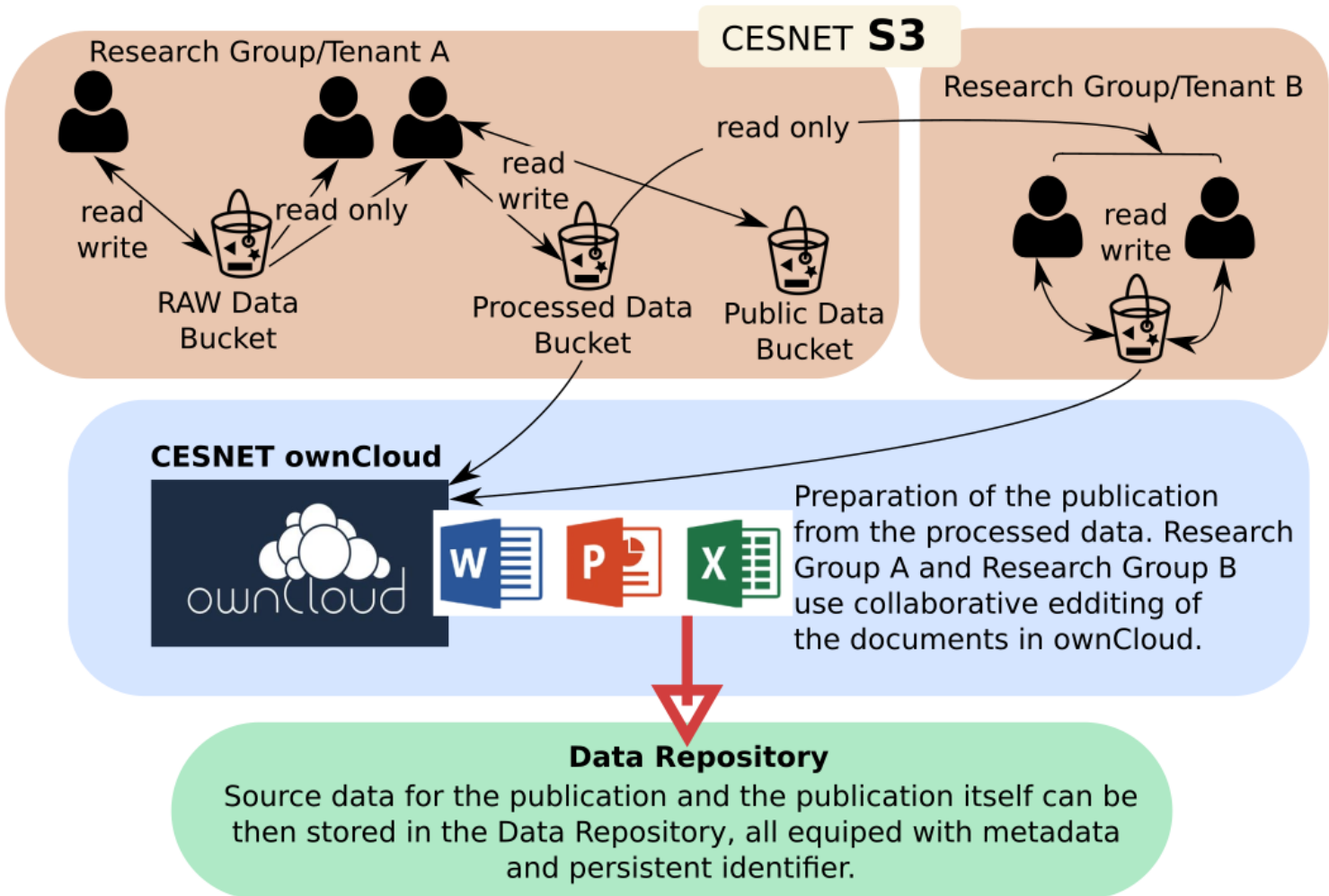


Popis případů užití RBD #2

Centralizované zálohování



Komplexní pracovní prostředí



Shrnutí

- S3 protokol pokryje velmi mnoho scénářů
- co nepokryje S3 může zvládnout RBD v kombinaci s nečím dalším

Otázky?

Popis případů užití RBD

Další případy užití budou ve formě reálných nasazení později v poslední přednášce

- Praktický příklad architektury pro zálohování ústavu AV ČR do CESNETu

AV CR a CESNET

- Primárně BTU AV CR, něco málo na PRF UK
- Potřeba zálohovat zhruba 100 zařízení
- Pracovní stanice, laptopy, servery, ruzná zařízení v laboratořích
- Stovky TB dat a stovky miliónů souborů

Požadavky na zálohovací SW

- Plné a inkrementální zálohy
- Řešení road-warrior modelu
- Zabezpečený přenos i přístup k zálohám
- Uživatel by měl mít možnost sám zkontrolovat stav záloh a případně si provést obnovu
- Windows, Linux, macOS

AV CR a CESNET

- Hrátky s rsyncem, duplicity, baculou
- Nakonec jsem zvolil BackupPC

BackupPC

- Open-source, GPL licence, napsané v Perlu
- **https://github.com/backuppc/backuppc**
- Webové rozhraní
- Deduplikace dat
- Komprese
- Má velmi rozsáhlou dokumentaci

BackupPC - webové rozhraní

- https/cgi
- Pro adminy i uživatele
- Stavy záloh, vyvolání nové nebo zrušení aktuální
- Procházení záloh a možnost jejich obnovy

BackupPC - záloha dat

- ICMP, true
- samba, tar přes ssh/rsh/nfs nebo rsync
- více disků, oddílů nebo adresářů
- include a exclude seznamy
- scheduler

BackupPC - obnova dat

- přímá obnova skrze smbclient, tar, nebo rsync/rsyncd
- nepřímá obnova přes tar nebo zip archív
- kvůli web rozhraní to může udělat i sám uživatel

BackupPC - Road-warrior

- Zařízení co jsou na síti jen občas a často i krátkou dobu
- DHCP a statické adresy
- WOL vychytávka

BackupPC - deduplikace dat a komprese

- Stejné soubory z jednoho nebo i více zařízení jsou uloženy pouze jednou
- Kompresi podléhají jen nové soubory (zatím neuložené do poolu)
- To vše šetří diskovou kapacitu, IO operace a CPU

BackupPC - nasazení

- RBD image s LUKS/dm-crypt
- s kolegou z CESNET jsme na obsluhu RBD imagů napsali Ansible role (dáme na GitHub)
- docker kontejnery (zafixovaná verze 4.x) přes docker-compose
- autentizace a autorizace přes LDAP BIOCEV

BackupPC - monitoring

- Icinga hlídá jen jestli běží weby na daných portech a místo na RBD
- Mailové notifikace - já i uživatelé
- Nejlepší monitoring jsou uživatelé
- Prometheus a Grafana pro performance metriky

BackupPC - co dále

- Definice nového zdroje pro S3 (rclone)
- Migrace z docker-compose do Kubernetes
- Dotáhnutí řádného monitoringu

Btrfs a ZFS na x způsobů

- Mějme kritický server mající stovky miliónů souborů
- Problém zálohovat na úrovni souborů
- Řešení je to dělat na úrovni bloků/snapshotu
- Send and receive (třeba přes btrbk)
- Další možný způsob, jak řešit ransomware apod.

Btrfs a ZFS na x způsobů

- Lokální snapshoty pro servery i pracovní stanice
- Snapshoty viditelné uživateli (připojené do cesty XYZ)
- Menší dopad na výkon systému (pozor na explozi snapshotů)
- Rychlé šahnutí na potřebná data
- Jedna s možností je snapper

Btrfs a ZFS na x způsobů

- Rozšíření diskového prostoru o vzdálený RBD image
- Opět možnost snapshotů
- Pozor na "živý přístup"a rychlost/latence/jitter po celé síti

S3 pre-signed requesty na x způsobů

- pre-signed requesty pro VM s výpočetním SW s licencí
- přílohy které jsou často i dost velké - OpenProject
- wrapper pro automatický upload po výpočtu

Shrnutí

- BackupPC s šifrovanym RBD image je dobrá volba
- Pro větší počet zařízení je dobré udělat více separátních instancí
- Pozor na "živý přístup"a rychlost/latence/jitter po celé síti
- Pozor na vícenásobné připojení RBD image

Otázky?