# Czy warto wdrożyć deduplikację danych w środowisku Tivoli Storage Manager?

**Paweł Mączka**
**Architekt Systemów Informatycznych**
**Pamięci Masowe i Archiwizacja Danych**
tel: 504273542
e-mail: p.maczka@infonet-projekt.com.pl

# Agenda

- Idea deduplikacji
- Trend czy realny zysk ?
- Deduplikacja w Tivoli Storage Manager
- Gdzie i kiedy deduplikować dane ?
- Monitoring środowiska TSM
- Pokaz implementacji i działania deduplikacji w TSM
- Podsumowanie

# Idea deduplikacji

- Deduplikacja danych jest procesem, który ma za zadanie porównanie i wyeliminowanie duplikatów danych składowanych na pamięciach masowych dla określonych systemów (systemy backupu/archiwizacji, serwery plików ).

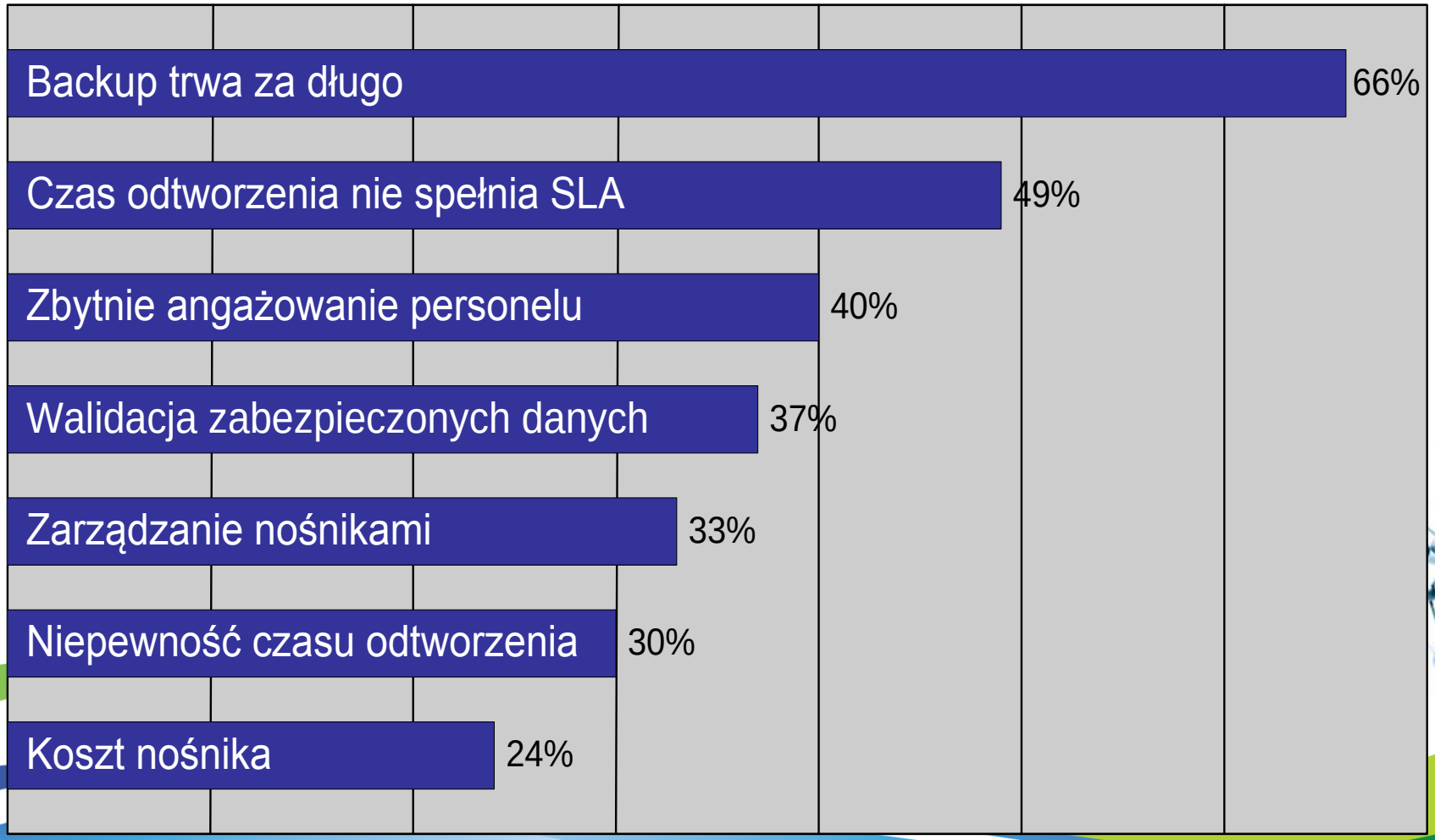
- Deduplikacja pozwala na znaczącą ich redukcję, co owocuje zmniejszeniem zapotrzebowania, kosztów poniesionych na pamięci masowe.

# Trend, czy realny zysk ?

infonet projekt

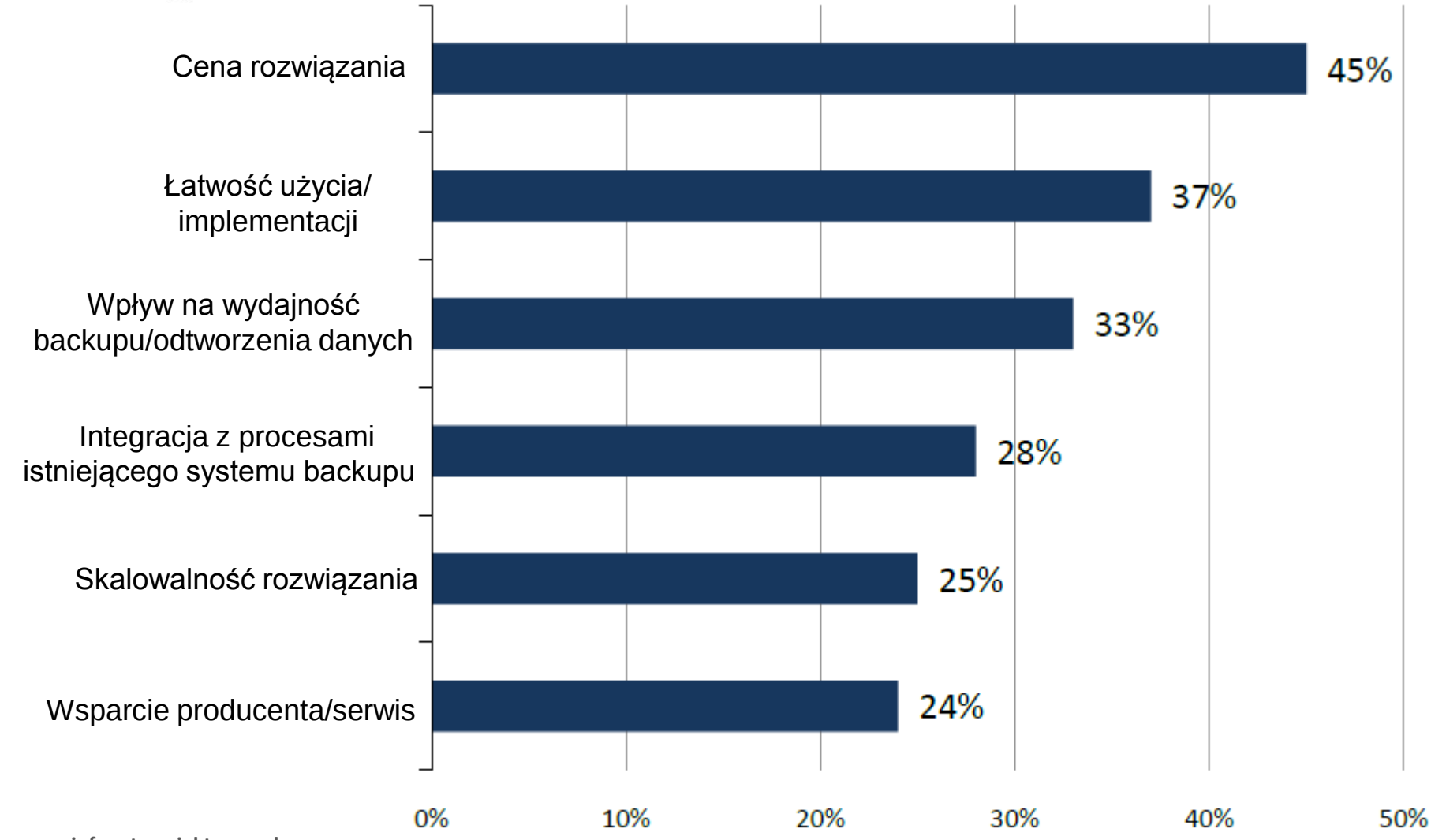
| | |
|---|---|
| Lepsze upakowanie danych | Składowanie większej ilości danych (x20 razy) na tej samej przestrzeni dyskowej |
| Optymalizacja procesu tworzenia kopii bezpieczeństwa | Dopasowanie do okna procesu |
| Zintegrowana deduplikacja danych | Zredukowanie kosztów związanych ze przestrzenią składowania danych, miejscem w serwerowni |
| Zarządzanie fizycznymi taśmami | Zredukowanie kosztów zarządzania |
| Prostsze zarządzanie taśmami offsite | Uproszczenie zagadnienia związanego z kopiami wynaszalnymi |
| Ochrona biur zdalnych | Wyeliminowanie kosztów i ryzyka związanego z taśmami, żonglowaniem nimi |

# Problemy dotykające większość firm na rynku



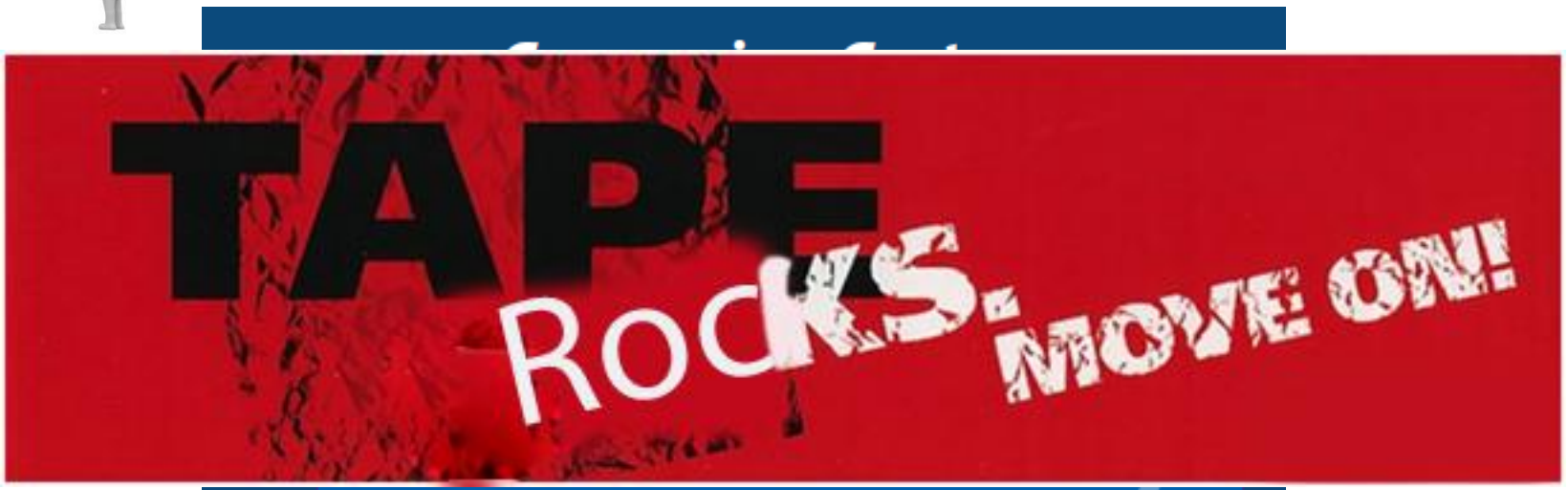

Percent of surveyed customers citing the problem. Source: ESG Enterprise Strategy Group

# Dlaczego decydujemy się na deduplikację ?

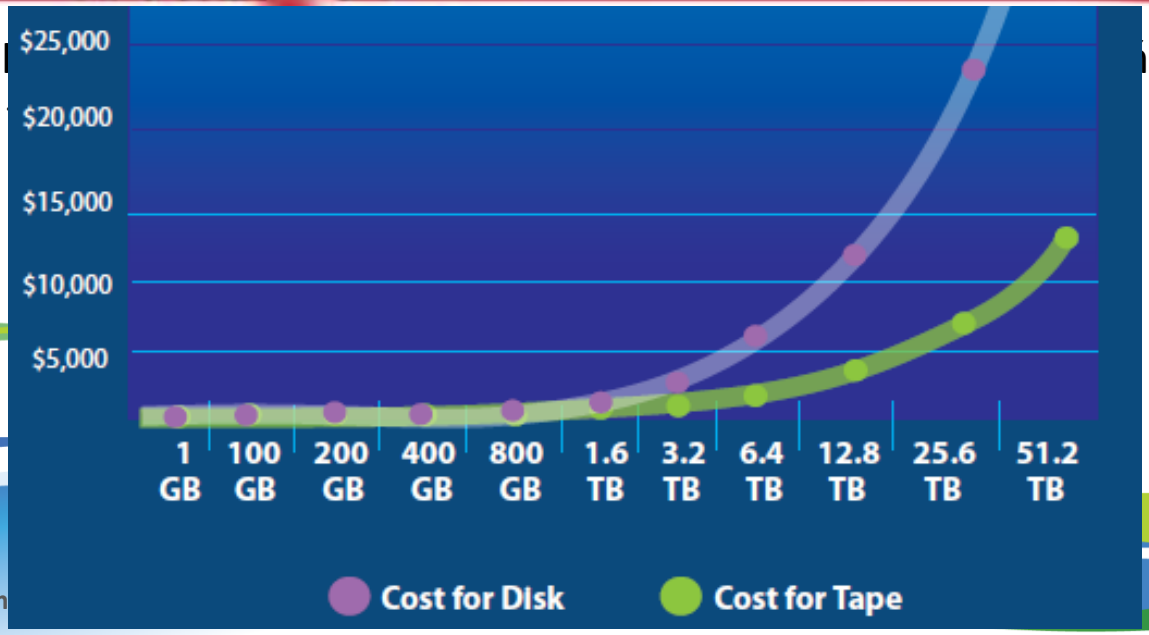



*Source: Enterprise Strategy Group, 2010.*

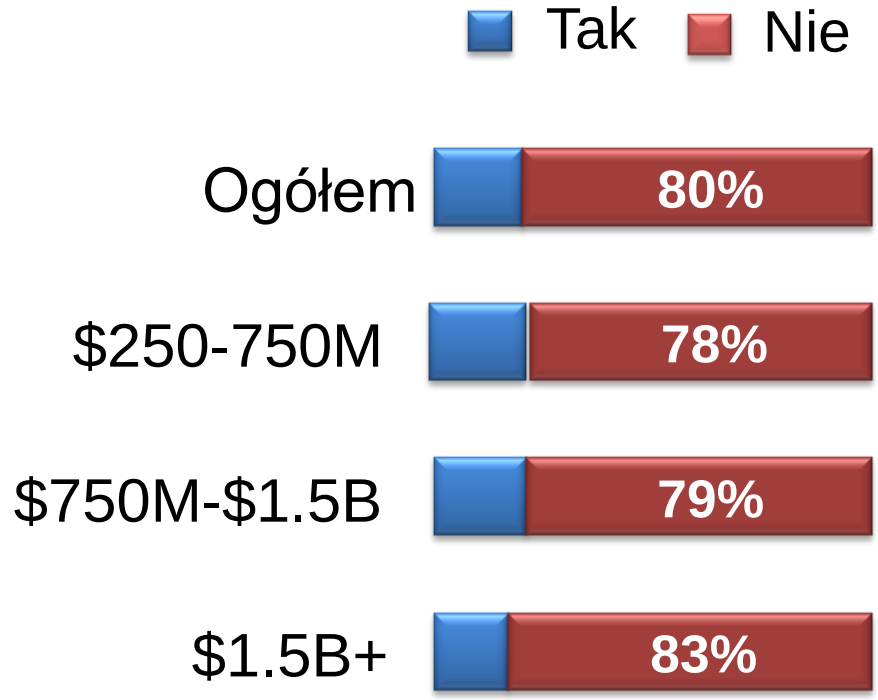
# Czy taśma przejdzie do lamusa ?




❑ I [illegible] czeniu ery backupu na [illegible]



# Czy wyeliminowanie taśm jest celem dla większości firm ?




Źródło: Forrester Research, 2008

# Backup/Archiwizacja
# D2D2T (Disk to Disk to Tape)

❑ **Złoty środek**. Metoda, tworząca hierarchię pamięci masowych, zrzucająca i składująca w odpowiedniej kolejności zabezpieczone dane najpierw na dysk, następnie na taśmie.


❑ Korzyści D2D2T:

- Szybki backup oraz odtworzenie
- Zmniejszenie okna backupowego
- Konsolidacja pamięci masowych
- Połączenie szybkiego dostępu na dyskach dla krytycznych systemów oraz aplikacji oraz długoterminowego przechowywania danych archiwalnych na taśmach

# Kiedy pojawiła się deduplikacja w TSM ?

er

e

upuje Tivoli

: Poprawiony
ści,
, wsparcie dla

TSM V4.1: backup systemów przenośnych i współdzielenie bibliotek

Tivoli Storage Manager (TSM) V3.7, następca ADSM

TSM V4.2: backup journal, po sieci SAN dla klientów, i SANergy

TSM 5.3: Zwiększenie wydajności i wsparcie dla dodatkowych aplikacji

TSM 5.4: poprawiona dostępności aplikacji i zarządzanie kluczami szyfrującymi

Tivoli CDP for Files: zabezpieczenie danych „w locie” dla użytkowników mobilnych

TSM FastBack: Continuous Data Protection dla serwerów Windows

TSM 5.5: podniesienie bezpieczeństwa danych, dotkowe wsparcie dla platform Windows, zabezpieczenie środowsk VMware i podniesienie skalowalności

TSM V6.2:
-Auto-deployment klienta TSM
-**Deduplikacja na kliencie**

TSM 6.1:
Nowa baza DB2, **dedupliakcja po stronie serwera TSM**, podniesiona wydajność, nowy silnik raportowania, Tivoli Monitoring

1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

# Deduplikacja w Tivoli Storage Manager

infonet projekt

- Występują 3 warianty, gdzie proces deduplikacji może się objawiać:

- źródło – serwer TSM

- cel – klienci TSM

-wirtualne biblioteki taśmowe - VTL

# Portfolio systemów backupu i archiwizacji danych związanych z deduplikacją

infonet projekt

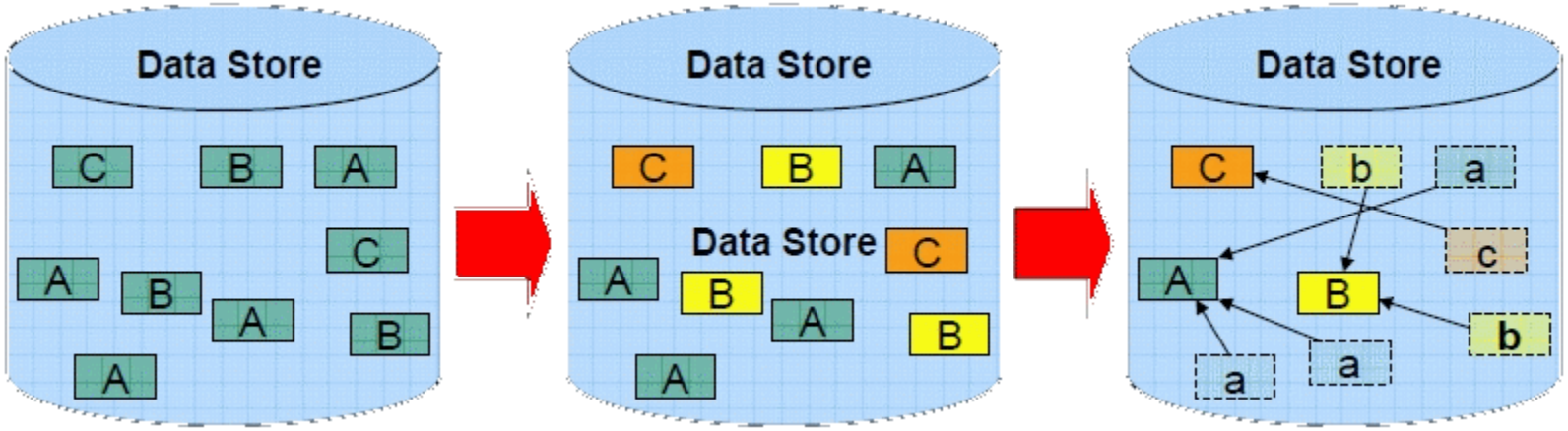
- ✓TSM wersja podstawowa
  - ✓Archiwizacja i odtworzenie z archiwum
  - ✓Zabezpieczenie i odtworzenie danych
  - ✓**Deduplikacja danych na kliencie oraz na serwerze**
- ✓TSM wersja rozszerzona
  - ✓Tivoli Disaster Recovery Manager (DRM)
  - ✓Wykorzystanie NDMP do zabezpieczenia urządzeń NAS
  - ✓Duże biblioteki
  - ✓**Deduplikacja danych na kliencie oraz serwerze**

- ➢TSM Fastback
  - ✓Rozwiązanie dla zdalnych lokalizacji
  - ✓Szybka instalacja, prosta konfiguracja
  - ✓CDP dla aplikacji
  - ✓ **Deduplikacja**

***Produkty serwerowe***

IBM TS7650G ProtecTIER Gateway

# Deduplikacja po stronie serwera TSM

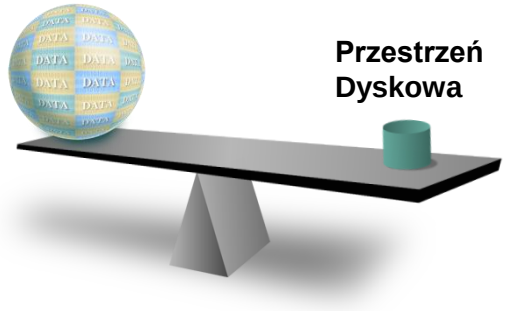



Dane znajdujące się, są dzielone na kawałki (chunks), oraz liczonę są ich sumy kontrolne

Wartości sygnatur są porównywane oraz identyfikowane jako zduplikowane

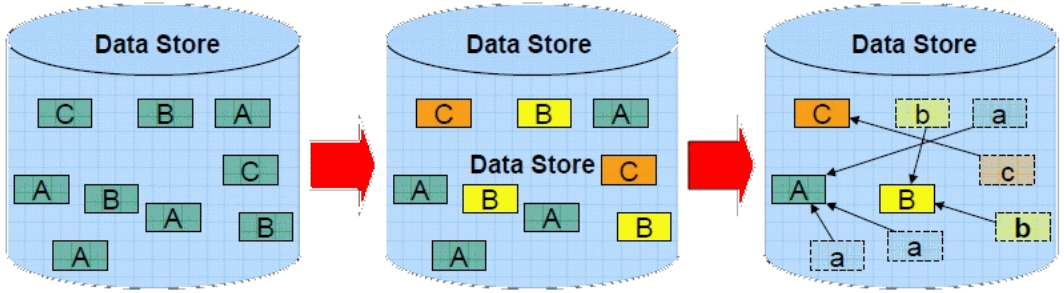
Zduplikowane kawałki (chunks) są zastępowane znacznikami do pojedyńczego kawałka

# Deduplikacja po stronie serwera TSM

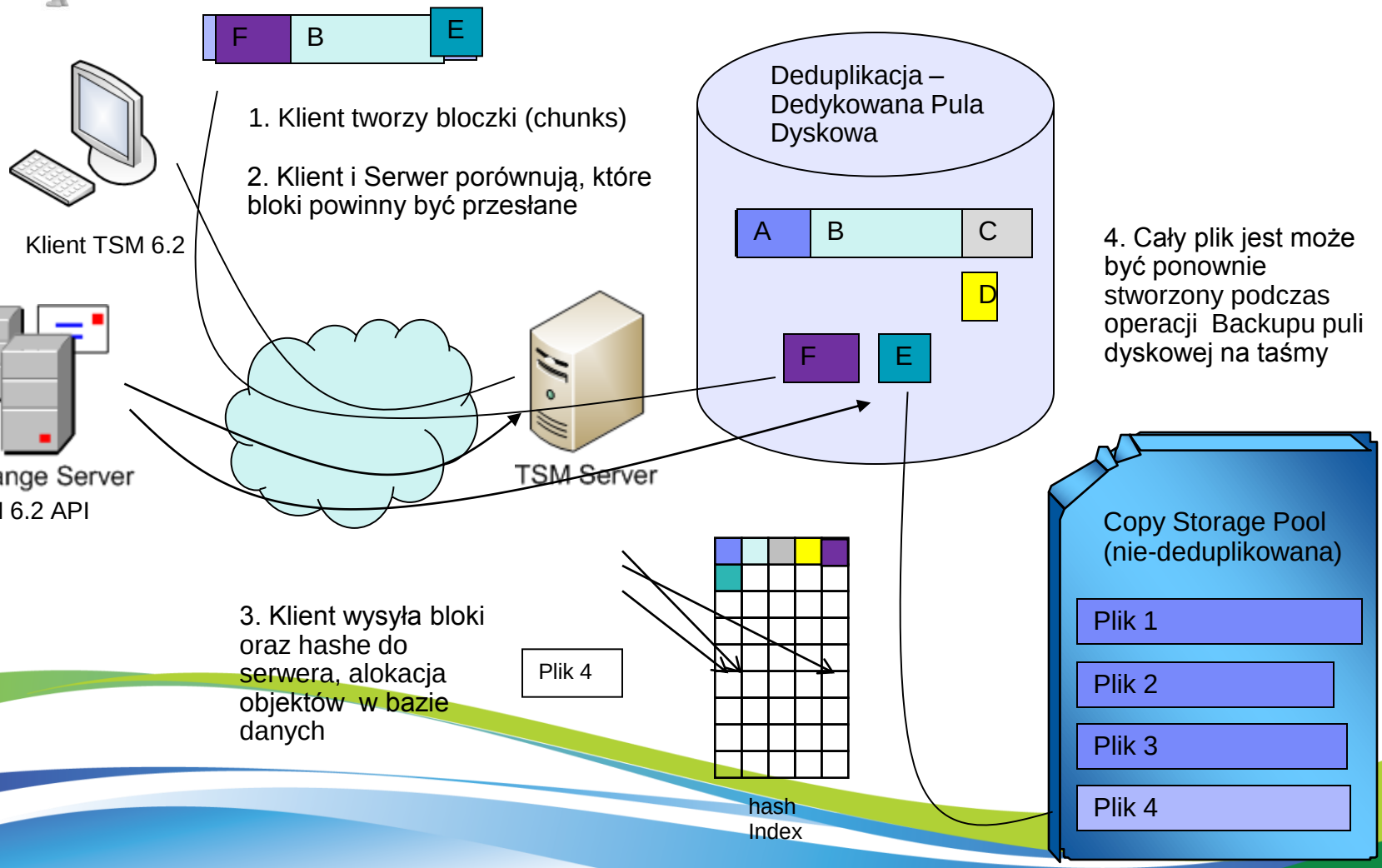



DEFINE / UPDATE STGPOOL stgpoolname
DEDUPlicate=No | Yes
IDENTIFYPRocess=nn

Identify DUPlicate sstgpoolname
DUration=mm
NUMPRocess=nn



# Deduplikacja po stronie klienta TSM



# Błędne porównanie (kolizja)

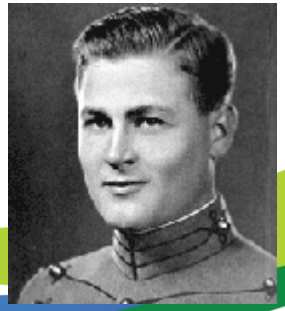
infonet projekt

- Istnieje prawdopodobieństwo, że dwa różne bloki danych po wykonaniu funkcji hash, dadzą ten sam wynik, co spowoduje tzw. kolizję i utratę unikatowego bloku

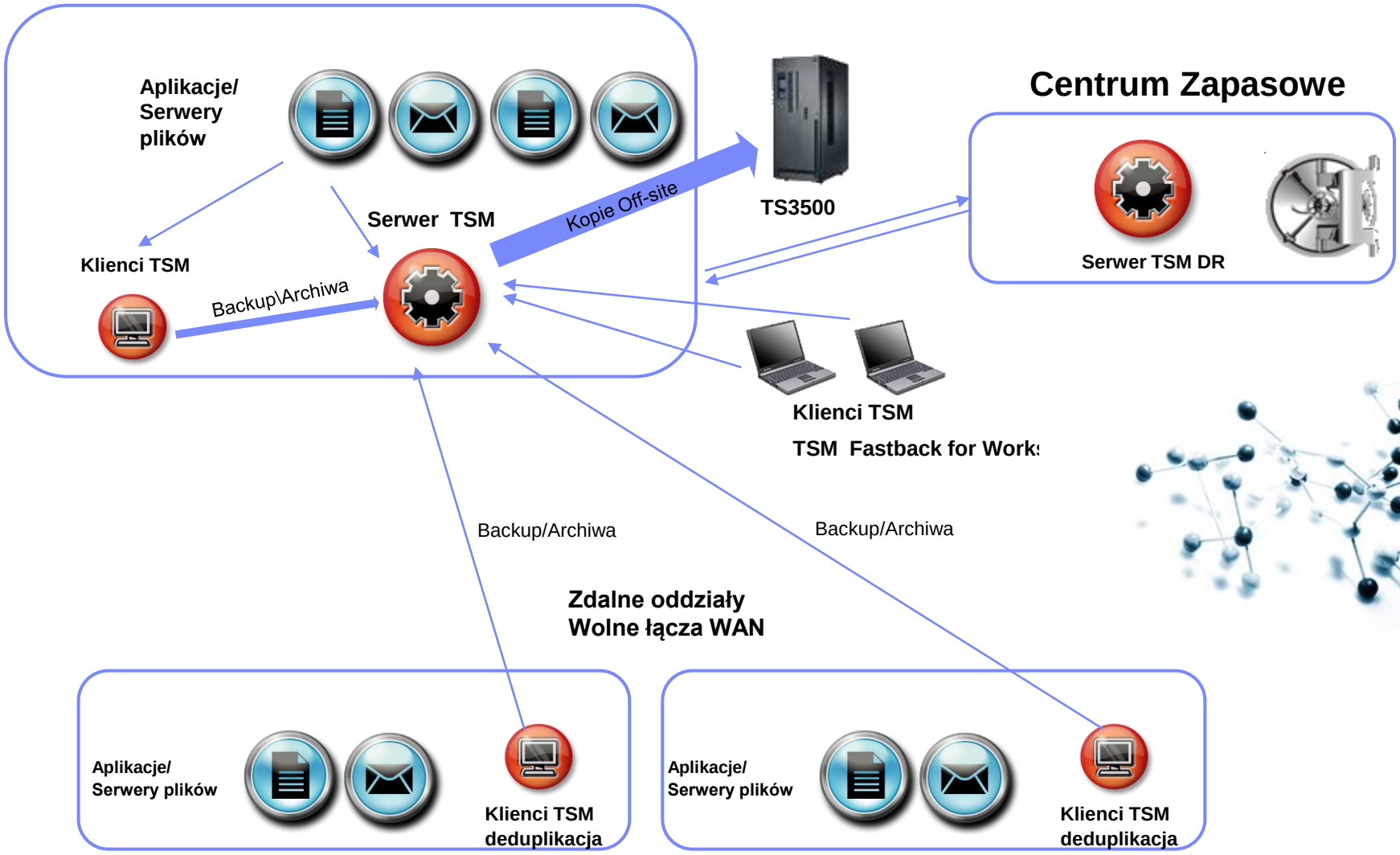
- Czy powinniśmy się martwić kolizjami ?
  - Algorytm użyty do porównywania bloków w procesie deduplikacji to SHA-1
  - Wystąpienie kolizji dla środowiska 4 PB, o bloku 4KB jest równe $0.5*10^{-28}$
  - Prawdopodobieństwo wystąpienia błędu na dysku to $10^{-14}$

"Jeśli deduplikacja ma jakąkolwiek możliwość zrobienia błędu, zrobi go"

**Kpt. Edward Murphy**

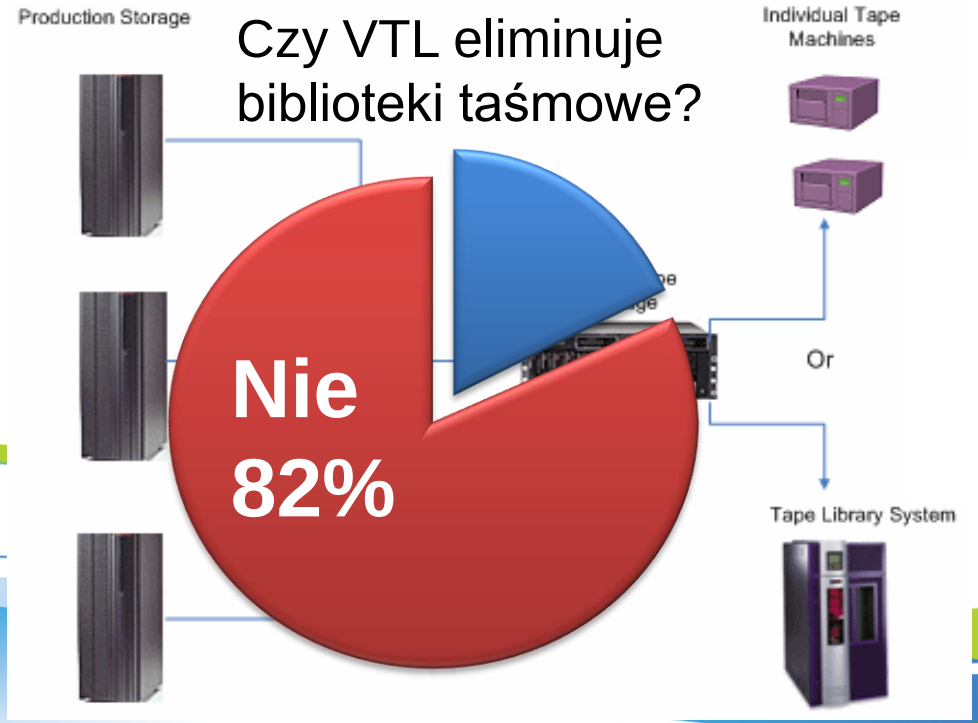


# Wydajny backup zdalnych oddziałów

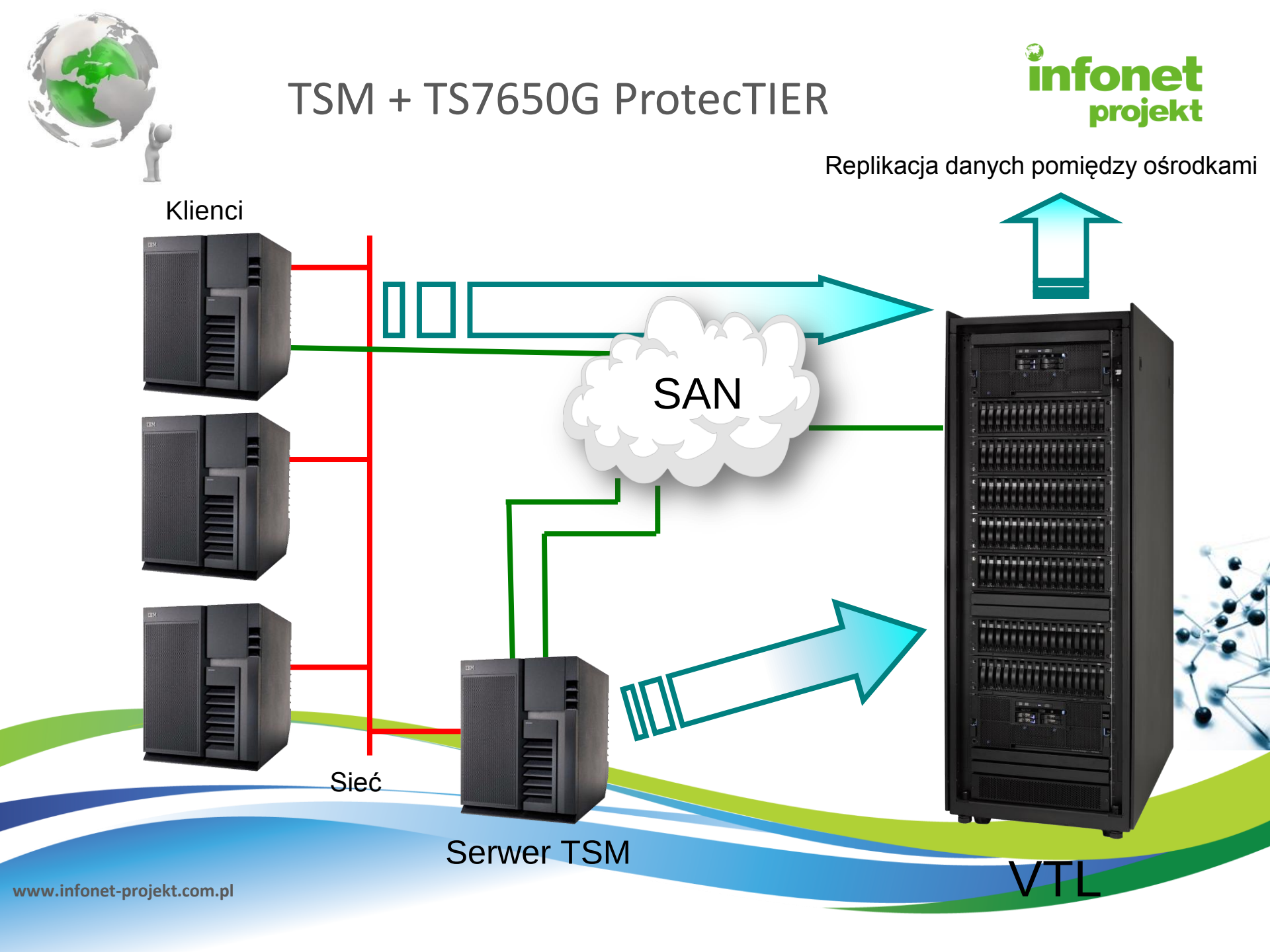


# Virtual Tape Library (VTL)

❑ **VTL** (Virtual Tape Library) – Wirtualna biblioteka taśmowa, w pełni potrafiąca zastąpić/uzupełnić fizyczną bibliotekę. VTL, emulując przedstawia się dla systemów backupu jako fizyczna biblioteka, pozwalająca na składowanie danych na wirtualnych taśmach, które są deduplikowane w locie, albo później poddane procesowi deduplikacji.



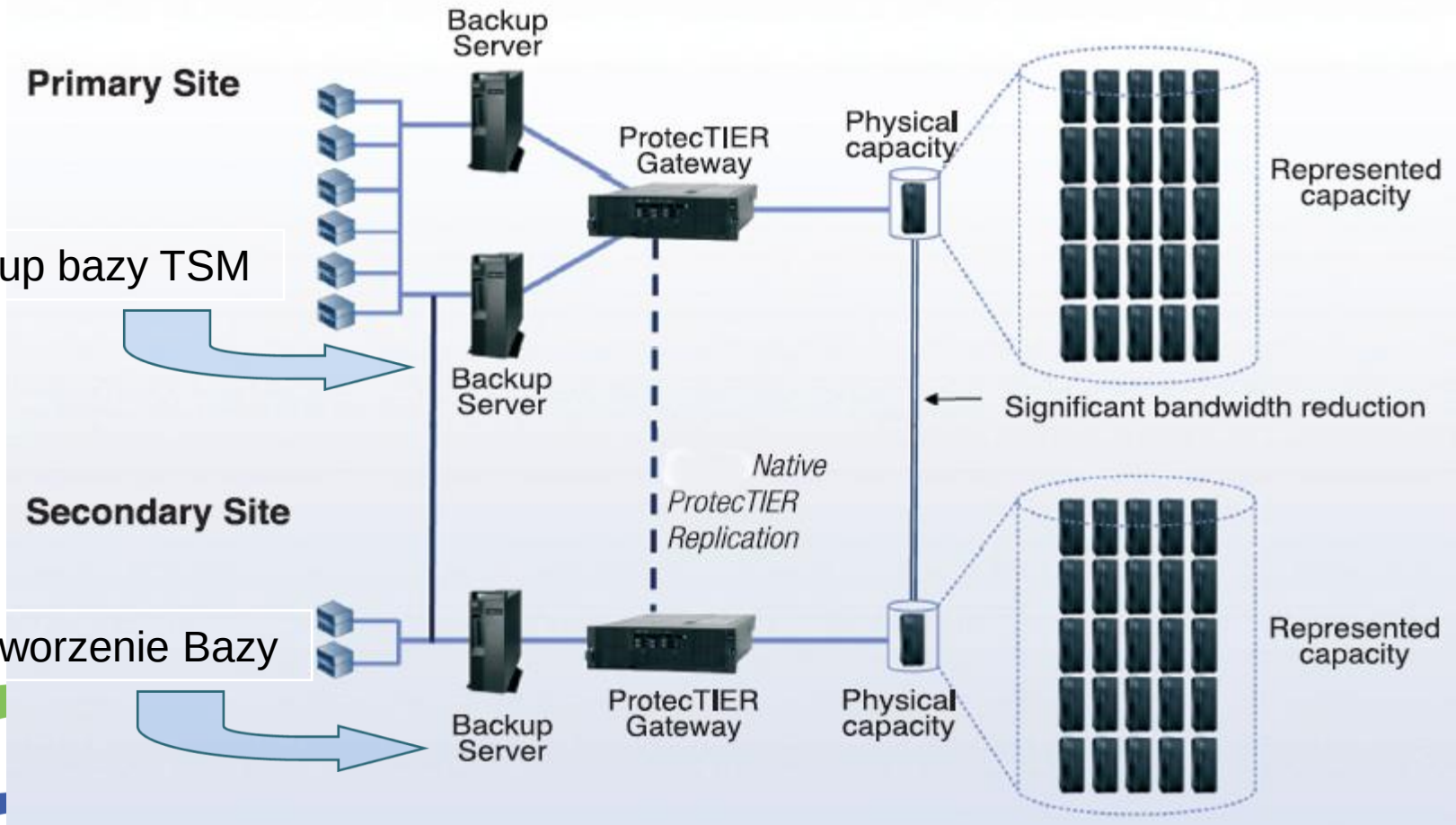
# TSM + TS7650G ProtecTIER



# Replikacja danych pomiędzy ośrodkami

Backup bazy TSM

Odtworzenie Bazy



# Gdzie deduplikować dane ?

infonet projekt

- Rozważmy dedupulikację na serwerze TSM jeśli:

  - Klient backupu ma wysoko obciążony procesor, lub jest to krytyczny system 24x7
  - Procesor oraz I/O dysku na serwerze TSM jest dostępne dla procesu porównania bloków
  - Klient nie działa w wersji 6.2
  - Chcemy deduplikować wszystkie dane (TSM API, wszystkie wtyczki)
  - Duże pliki na mogą się wolniej odtwarzać, jeśli mamy deduplikację po stronie klienta

- Rozważmy dedupulikację na kliencie TSM jeśli:

  -Klient backupu ma nisko obciążony procesor
  - Jest realizowane zdalne, wolne łącze pomiędzy klientem a serwerem backupu
  - Ta metoda wydaje się bardziej skalowalna. Możemy dodać większą ilość klientów, nie przejmując się wpływem deduplikacji na obciążenie serwera TSM
  - Posiadamy serwery, które współdzielą te same dane
  - Posiadamy duża liczbę systemów Windows – backup systemstate

# Gdzie deduplikować dane ?

infonet projekt

- A może jednak zarówno na kliencie i serwerze, w zależności od obciążenia ?

  - W weekendy, kiedy sieć jest mniej obciążona, użyjmy deduplikacji po stronie serwera TSM
  - W trakcie tygodnia, gdzie czas odpowiedzi sieci jest kluczowy, użyjmy deduplikacji po stronie klienta TSM
  - Kontrola poprzez zastosowanie makra – update node ”XXX” deduplication=serveronly

- Może jednak w ogóle nie stosować ?

  - Odtworzenie z deduplikowanych danych może zająć więcej czasu. Plik może być rozrzucony po wielu wolumenach, co owocuje zwiększeniem odwołań I/O do zdeduplikowanej puli.
  - Odtworzenie z zdeduplikowanej puli owocuje zwiększonym I/O bazy danych TSM
  - Kluczowe, krytyczne systemy powinny lądować na puli aktywnej, lub standardowej nie zdeduplikowanej

# Gdzie deduplikować dane ?

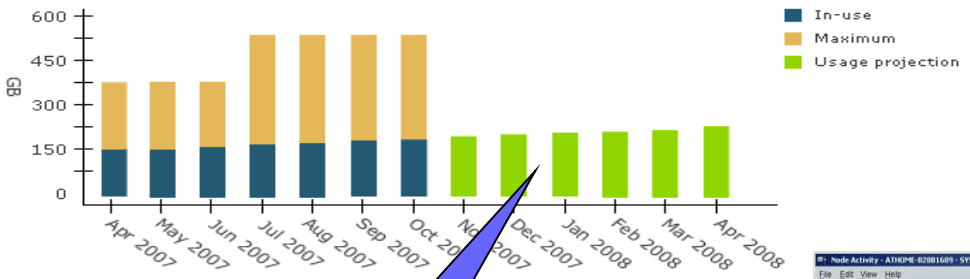
- Deduplikacja VTL jest realizowana albo locie, albo w czasie zadanym przez administratora na dedykowanych sprzęcie. Proces ten w w ogóle nie wpływa na obciążenie serwerów backupu.
- Pozwala na backup po sieci SAN na dyski.
- Najbardziej optymalnie wydaje się połączenie deduplikacji po stronie klienta TSM dla mniej krytycznych danych, z VTL podłączonym do serwera TSM oferującym deduplikację, mechanizmy replikacji, DR.
- Należy pamiętać, że deduplikacja na serwerze jak i na kliencie od wersji 6.2 (maj 2010) jest dostępna w Standard Edition

# Monitoring środowiska TSM

Tivoli. IBM

## Storage Capacity Projection

Start Date 2007-04 End Date 2008-04




| Client Node Name | Maximum | In-use (GB) |
|---|---|---|
| FILESRV01 | | 100 |
| FILESRV02 | | 160 |
| DB2_SRV_FIN | | 160 |
| DB2_SRV_DEV | | 120 |

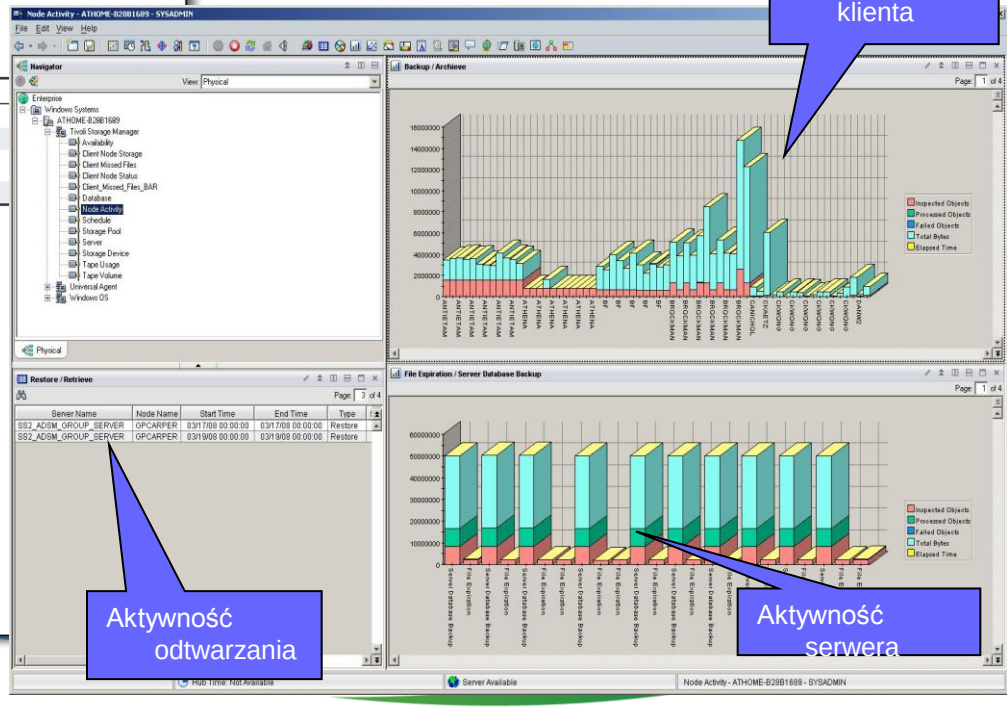
2007-10-27 23:02 EST

Przewidywanie wymagań

*W pełni modyfikowalne raporty*

*Zawarte w licencji podstawowej – bez dodatkowych licencji*

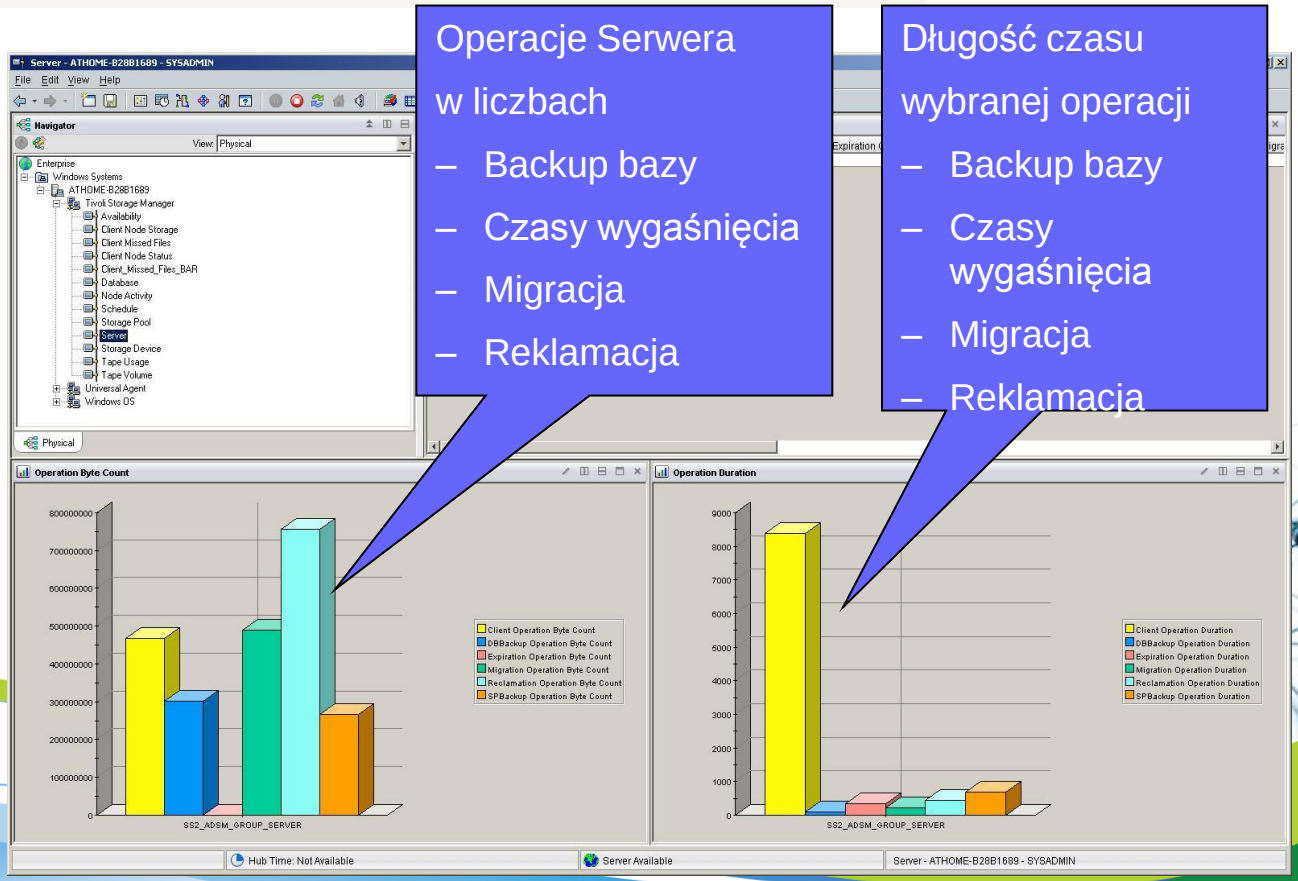
*W pełni modyfikowalny wygląd*



# Monitoring środowiska TSM

***Modyfikowalny wygląd, dostarcza informacji w postaci graficznej na temat statusu i działania systemu TSM***

- ✓ Scheduled Client Activity
- ✓ Scheduled Server Activity
- ✓ Client Current Activity
- ✓ Server Current Activity
- ✓ Current Errors
- ✓ TSM Database Status
- ✓ Tape Device Status
- ✓ Storage Pool Status
- ✓ Client Backup Status



infonet projekt

# Pokaz implementacji i działania deduplikacji w TSM


Pytania

infonet projekt

# Dziękuję za Uwagę

**Paweł Mączka**
**Architekt Systemów Informatycznych**
**Pamięci Masowe i Archiwizacja Danych**
tel: 504273542
e-mail: p.maczka@infonet-projekt.com.pl