Rozproszone składowanie i przetwarzanie danych dostępność, bezpieczeństwo, poufność

Michał Białoskórski
CI TASK

Infobazy 2014

Big Data

- Zwiększająca się moc obliczeniowa procesorów
  - Zwiększa się liczba operacji
  - Zwiększa się ilość przetwarzanych danych
  - Zmniejsza ilość kWh na liczbę operacji
  - Coraz częściej klastry – z uwagi na cene
- Coraz większe zapotrzebowanie na składowanie danych

Big Storage

Coraz większe potrzeby składowania danych

- Wciąż dyski 7200 obr/min
  - Szybkość transferu rzędu 100MB/s
- Dobre dyski typu Flash bardzo drogie
- Duże macierze z odpowiednią szybkością - bardzo drogie

- Problem z odpowiednim zabezpieczeniem danych

Składowanie rozproszone

Klaster (zespół) zasobów dyskowych

- Złożony z
  - Serwerów udostępniających
  - Zasobów dyskowych
- Serwery dużo tańsze od zasobów dyskowych
- Możliwość zabezpieczenia przez kopie danych na innym zasobie dyskowym

Składowanie rozproszone

- Skalowalne urządzenia
  - Łatwa rozbudowa o dodatkowe miejsce
  - Łatwe zwiększenie szybkości transferu
  - Łatwe zwiększenie ilości operacji I/O
- Bezpieczeństwo
  - Replikacja danych
  - Większa odporność na awarię

Jak - oprogramowanie?

- Darmowe
  - LUSTRE
  - CEPH
  - LizardFS/MooseFS
  - Hadoop
- Komercyjne
  - SCALITY
  - IBM GPFS
  - Wsparcie techniczne do darmowych

Przetwarzanie w chmurze

- Brak zapewnienia poufności
- Potrzeba zabezpieczenia po stronie klienta
- Chmury tylko od zaufanych dostawców
- Dobre przyłącze do internetu

- Brak odporności na awarię łącza

Bezpieczeństwo i poufność

Bezpieczeństwo

- Zagrożenie utraty danych
- Zagrożenie kradzieży danych

Poufność

- Zagrożenie udostępnienia zawartości danych

Bezpieczeństwo

- Odpowiedni wybór oprogramowania
- Odpowiedni wybór nośników danych
- Zabezpieczenie danych
  - Kopia zapasowa
  - Zdalna replika danych
  - Kontrola spójności danych

Poufność danych

- Analiza zagrożeń
- Niwelacja zagrożeń
  - Szyfrowany dostęp do danych
  - Szyfrowanie przynajmniej składowanych metadanych
  - Oddzielenie metadanych od danych
- Dobór odpowiedniego operatora danych

Rozwiązania w CI TASK

Specyfika CI TASK

- Komputery Dużych Mocy obliczeniowych (KDM)
  - Duża moc obliczeniowa
  - Duże środowisko składowania danych
  - Archiwizacja wyników
- Usługa Powszechnej Archiwizacji Danych (PLATON-U4)
  - Duża ilość składowanych danych
  - Bezpieczeństwo – zdalne repliki danych

KDM

- Klastry obliczeniowe
  - Galera Plus (50 TFLOPS) 2008/2012r.
  - Nowe zasoby (1200 TFLOPS)
- Rozproszony system plików
  - RORO – 2008r.
  - RORO-2 - 2014r.
- Magazyn danych
  - HSM

RORO (2008)

Dane:

- System plików LUSTRE
- 12 serwerów dyskowych (576 dysków 2TB)
- 2 serwery meta-danych
- 1 macierz **3par T800** (140 dysków 15kRPM)

Szybkość zapisu 10GB/s

Brak zabezpieczenie przed awarią serwera dyskowego

RORO-2 (2014)

Dane

- System plików LUSTRE
- 4 serwery danych
- 2 serwery meta-danych
- 1 macierz **Hitachi HUS-150** (108 dysków 10kRPM)
- 3 macierze **Hitachi HUS-VM** (384 dyski 3TB)

Prędkość zapisu 15 GB/s

Pełne zabezpieczenie przed awariami sprzętu

ci task

Magazyn danych

Podłączony do serwera udostępniania danych

- Oprogramowanie IBM TSM
- Macierz IBM DS5300
- Biblioteka taśmowa IBM TS3500
  - 2000 taśm LTO5 z szyfrowaniem
  - 16 napędów taśmowych

Bezpieczeństwo

Kopia zapasowa
- Tylko katalogi domowe (20TB)
- Brak dla katalogów roboczych (setki TB)

Sprzęt
- RAID-6 na dyskach
- Klaster wysokiej dostępności (RORO-2)

Dostęp
- Szyfrowane kanały dostępu
- Dwupoziomowy dostęp do zasobów

Dziękuję za uwagę