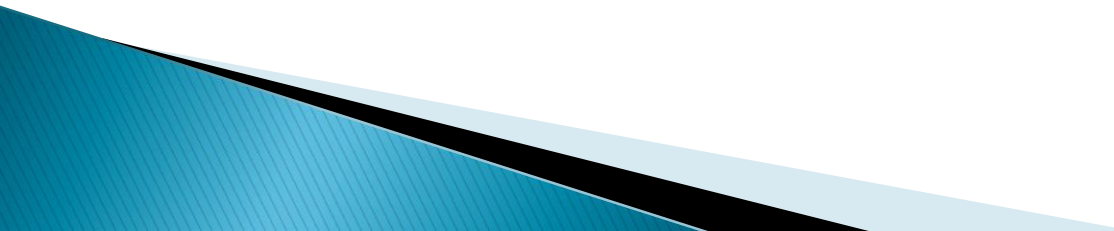


Komprese XML souborů

Pavel Hruška

Obsah prezentace

- ▶ Stručně o XML (výhody, nevýhody)
 - ▶ Proč komprimovat XML?
 - ▶ Metody komprese XML
 - ▶ XMill
 - Představení, princip, výsledky
 - Analýza XML (možná úskalí)
 - ▶ Optimalizace komprese XML
 - Přeskládání kontejnerů (XMill)
 - Přeskládání celých XML souborů
 - ▶ Závěr
- 

Stručně o XML

- ▶ XML = eXtensible Markup Language
- ▶ Otevřený značkovací jazyk
- ▶ Standardizovaný konsorciem W3C
- ▶ Sada pravidel pro tvorbu vlastních jazyků
- ▶ Univerzálnost + široká podpora => populární jazyk
 - B2B, konfigurační soubory, propojení systémů, univerzální datový formát, ...

Proč komprimovat XML?

- ▶ XML platí za své výhody výřečnosti
 - V porovnání s nativními binárními formáty větší datová náročnost
- ▶ Problémy?
 - Uložení dat (archivace)
 - Přenos (internet, WAN, ...)
 - Práce s XML (analýza zatěžuje procesor, paměť)
- ▶ Řešení pomocí komprese XML
 - Snížení datové náročnosti
 - Méně náročné zpracování
(algoritmy komprese s podporou dotazování)

Metody komprese XML

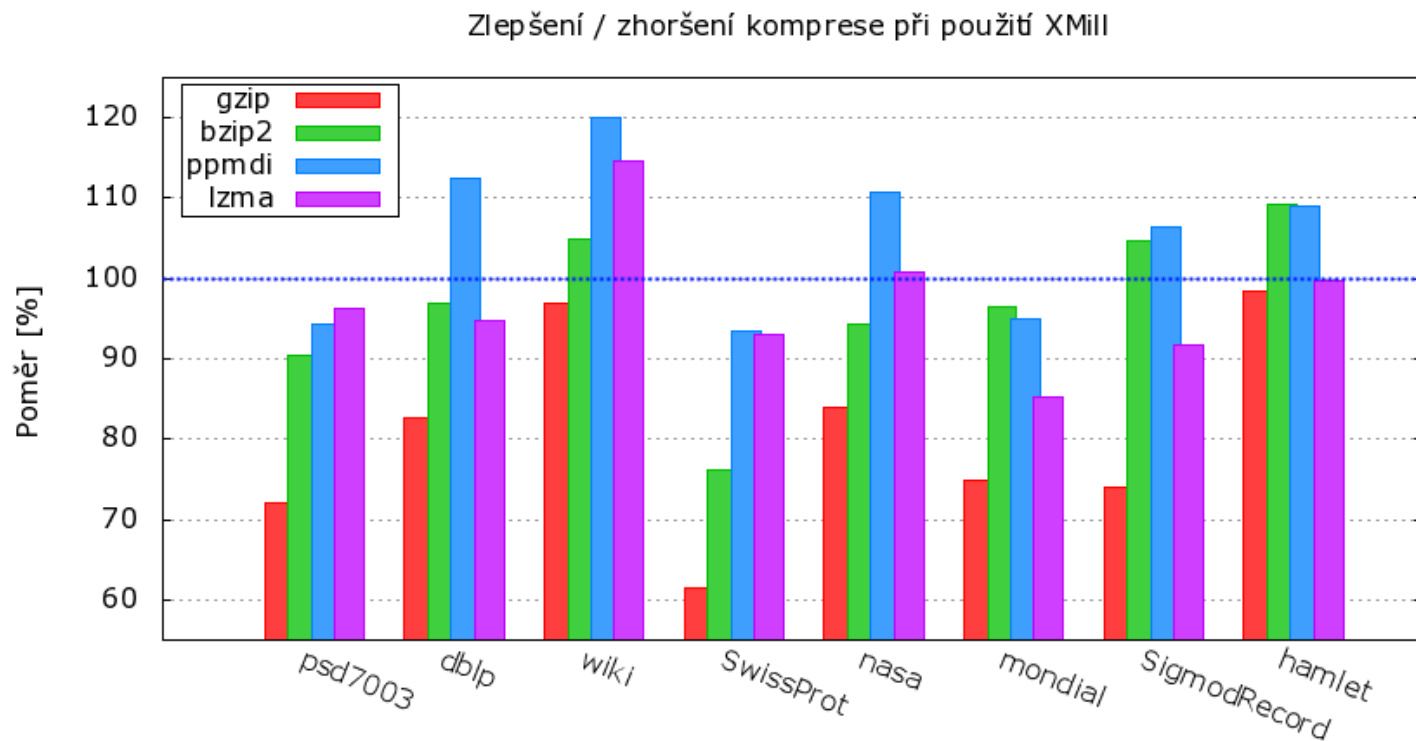
- ▶ Komprese XML jako textu
 - Pomocí běžných kompresních algoritmů
 - Metody: Deflate, BZip2, PPM, LZMA, ...
 - Programy: 7-Zip, WinRAR, ...
- ▶ Komprese XML-aware kompresory
 - Bez podpory dotazování
 - Využívají strukturu XML
 - Připravují data pro efektivnější kompresi běžnými kompresory => finální komprese: Deflate, BZip2, ...
 - S podporou dotazování
 - Umožňují pracovat s komprimovanými XML daty (dotazy, modifikace)

XMill

- ▶ XML-aware kompresor (bez dotazování)
- ▶ Odděluje strukturu od dat
 - Strukturu kóduje pomocí slovníku
- ▶ Využívá sémantiku v XML datech
 - Značka dává význam v ní uzavřenému obsahu
- ▶ Data seskupuje dle sémantiky do kontejnerů
- ▶ Finální komprese běžnými algoritmy
 - Standardní podpora gzip, bzip2 (knihovna *zlib*)
- ▶ Podpora sémantických kompresorů
 - Kódování celých čísel, výčtů, sekvencí (IP adresa) (specifikuje ručně uživatel)

Kompresa XMill v praxi (1 / 2)

- Srovnání kompresních algoritmů běžnou kompresí (text) a s využitím XMill



Kompresa XMill v praxi (2 / 2)

► Shrnutí experimentů:

- V průměru zlepšení komprese:
 - Deflate ~ -19%
 - BZip2 ~ -5%
- V průměru stejné výsledky:
 - LZMA
- V průměru zhoršení:
 - PPMdI ~ +7%

► Závěr:

- XMill byl navržen pro Deflate a BZip2.
- Moderní algoritmy (LZMA, PPM) dosahují v průměru stejných nebo lepších výsledků

Analýza XML a XMill (1 / 3)

- ▶ Problém analýzy XML a seskupování:
ne vždy má značka sémantický význam
- ▶ Příklad:
`<text>Pozor: <b>Nebezpečné</b></text>`
 - Značka `<b>` pouze formátuje, nemá sémantický význam.
 - Současné kompresory to neřeší => dochází k chaotickému seskupování a kouskování dat => horší komprese.
- ▶ Konkrétní příklad – soubor wiki.xml

Analýza XML a XMill (2 / 3)

```
<wiki>
<article>
<name id="1000003">Rosie Beaton</name>

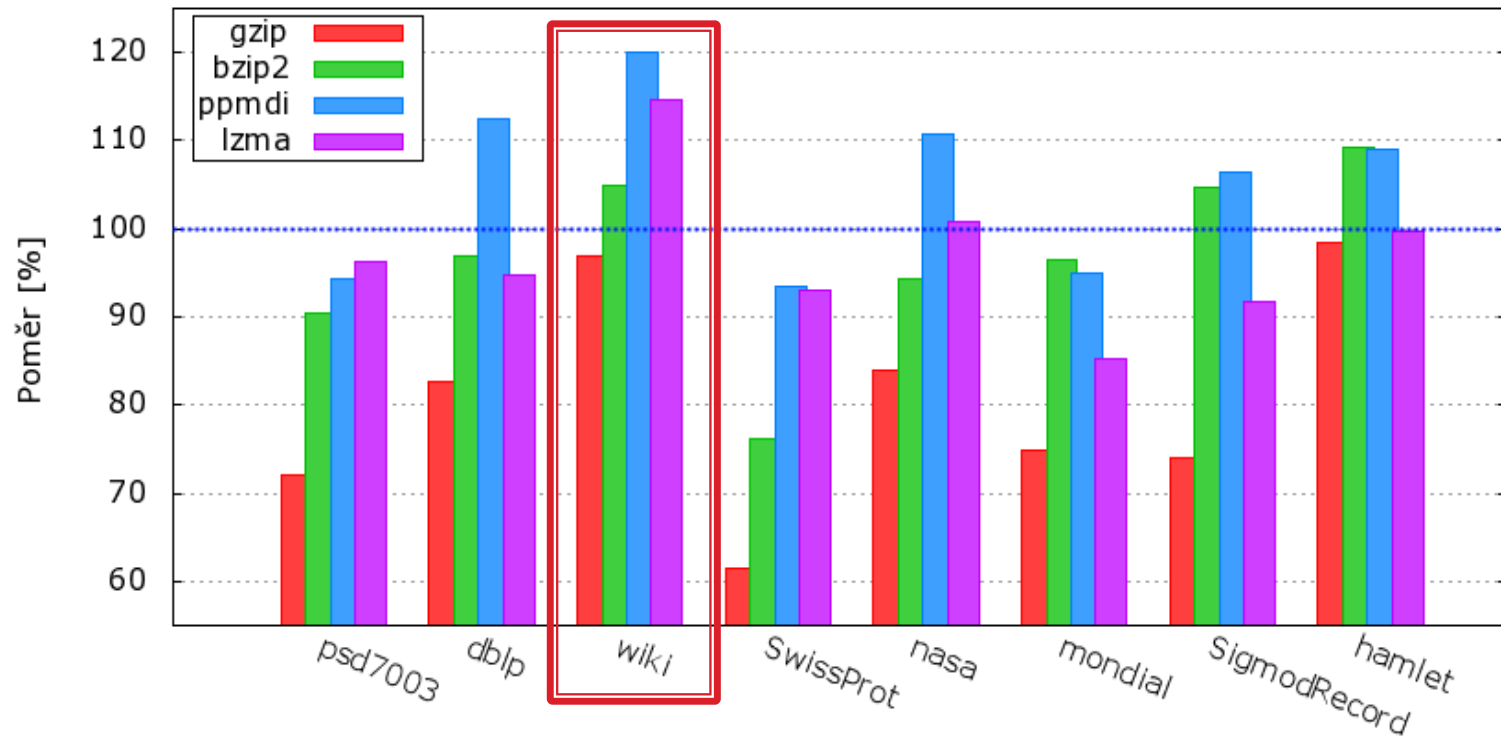
<conversionwarning>0</conversionwarning><body>

<emph3>
Rosemary (Rosie) Beaton
</emph3> (born
<unknownlink src="29 May">
29 May
</unknownlink>) is an
<collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type=
Australian
</collectionlink> radio announcer, best known for her work at Australi
<collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type=
radio station
</collectionlink><collectionlink xmlns:xlink="http://www.w3.org/1999/x
Triple J
</collectionlink>.
<section>
<title>Radio career </title><p>
Rosie Beaton is currently the host of
<collectionlink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type=
Triple J
```

Analýza XML a XMill (3 / 3)

- ▶ wiki.xml – obsahuje články s formátováním

Zlepšení / zhoršení komprese při použití XMill



Optimalizace komprese XML

s využitím přeskládání XML

Optimalizace komprese XML

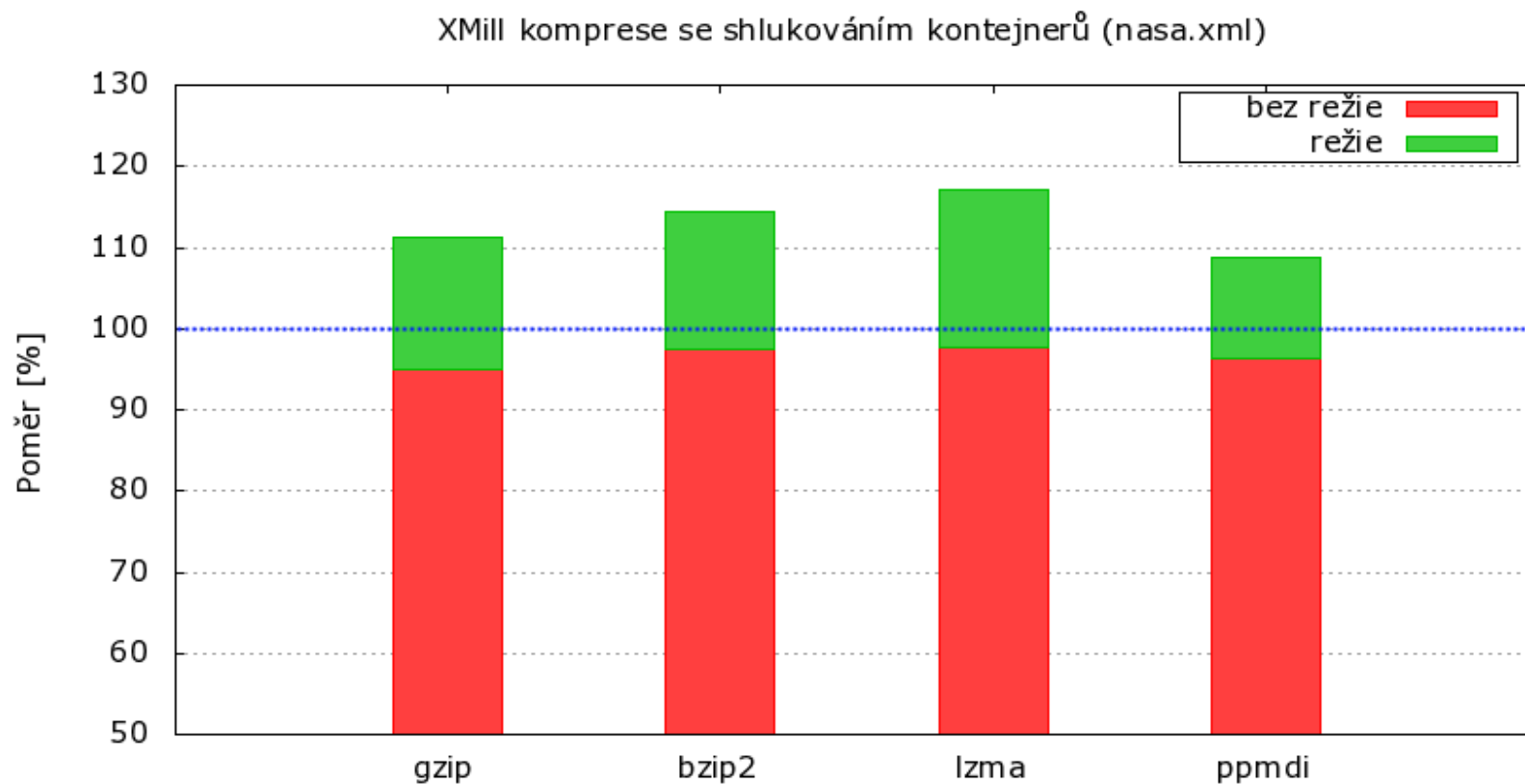
- ▶ Provedené experimenty
 - Přeskládání dat v kontejnerech XMill
 - Přeskládání celých XML souborů

Přeskládání kontejnerů (1 / 2)

- ▶ Přeskládání prvků kontejneru před kompresí
- ▶ Shlukování kontejnerů zlepšuje jejich jednotlivou kompresi
- ▶ Ale jsou zde problémy:
 - Časově náročný algoritmus shlukování (komprese 25MB XML dat více jak 3 hodiny)
 - Architektura XMill vyžaduje uložení původního uspořádání prvků => režie navíc
- ▶ Celkově neefektivní
 - Pomalé
 - Výsledná velikost horší než bez shlukování

Přeskládání kontejnerů (2/2)

► Výsledek komprese nasa.xml



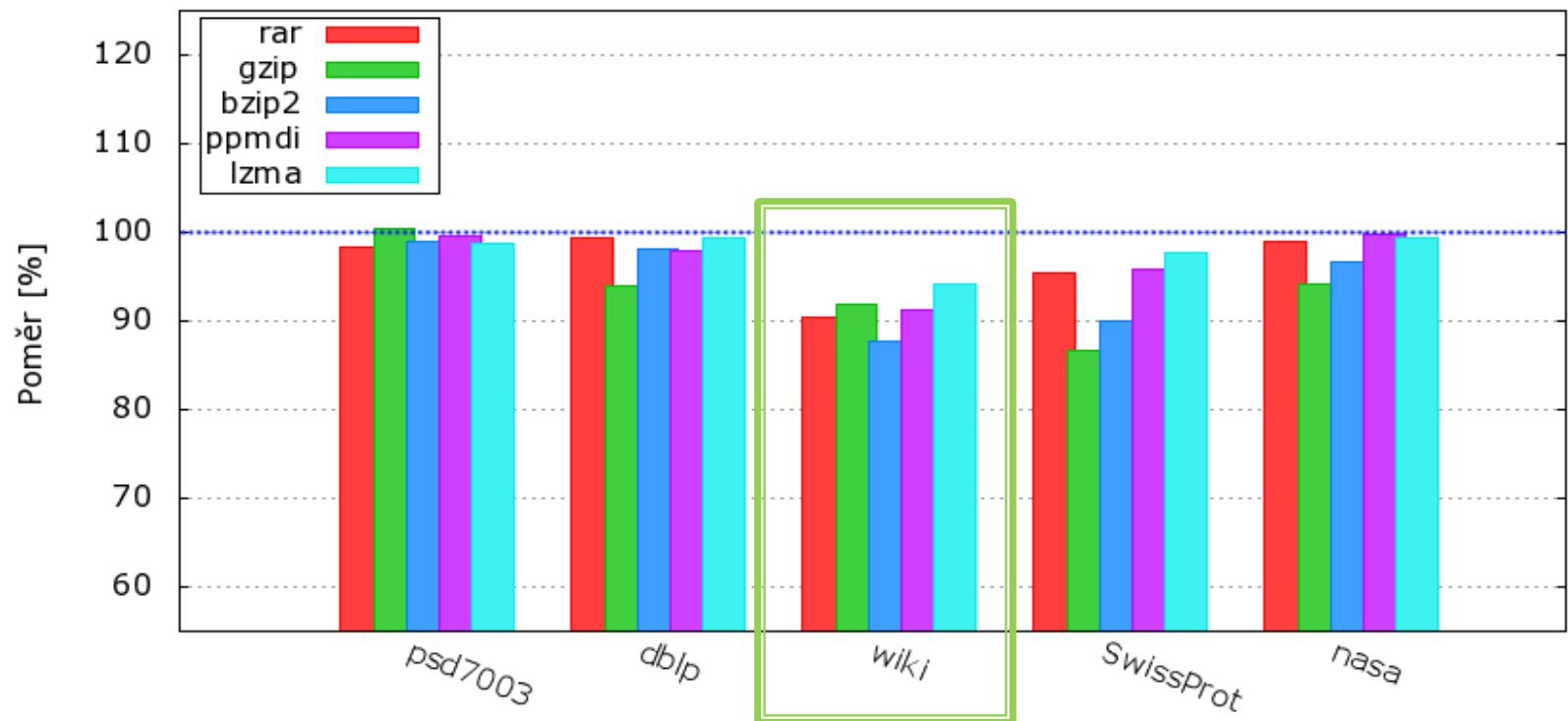
Přeskládání celých XML (1 / 4)

- ▶ Přeskládání XML
 - Není vhodné pro všechny XML dokumenty (nesmí zakazovat například schéma)
- ▶ Použitelné pro kompresi XML jako textu i pro kompresi pomocí XMill
- ▶ Použité metody shlukování jsou
 - časově přijatelné (konkrétně cca 30–60 minut)
 - ale paměťově náročné (testovací stroj 16GB RAM)

Přeskládání celých XML (2/4)

► Komprese shlukovaných XML jako textu

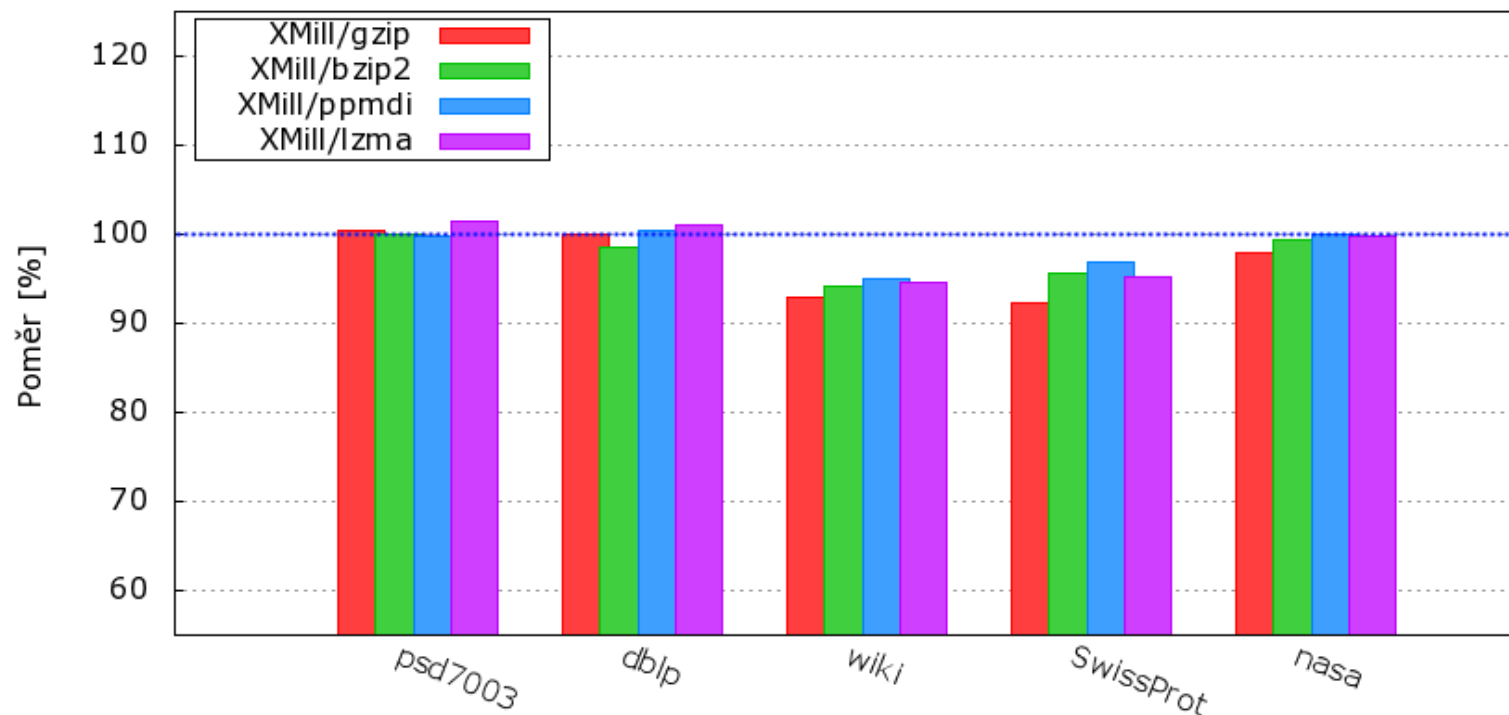
Srovnání běžné komprese po provedení shlukování celých XML souborů



Přeskládání celých XML (3 / 4)

► Komprese shlukovaných XML pomocí XMill

Srovnání XMill komprese po provedení shlukování celých XML souborů



Přeskládání celých XML (4/4)

- ▶ Shrnutí shlukování celých XML
 - Dobré výsledky při kompresi XML jako textu, tak při kompresi XML pomocí XMill
 - Maximálně mírné zhoršení (napříč všemi algoritmy)
 - Účinné i na soubory, u kterých není úspěšná strukturální komprese (XMill)

Závěr

- ▶ Komprese XML
 - Běžné textové kompresory
 - Specializované XML-aware kompresory
- ▶ Možná úskalí komprese XML
 - Důležitým prvek komprese je analýza XML (oddělení struktury od dat)
- ▶ Možnosti zlepšení komprese XML
 - Pomocí přeskládání dat
 - Přeskládání kontejnerů (XMill) – možná cesta, nutno dořešit rychlost algoritmu a ukládání režie
 - Přeskládání celých XML – dobré výsledky

Konec prezentace

Děkuji za pozornost

Dotazy?