Prezentace diplomové práce

Aplikace kompresní metody DCA

Bc. Jan Skalický

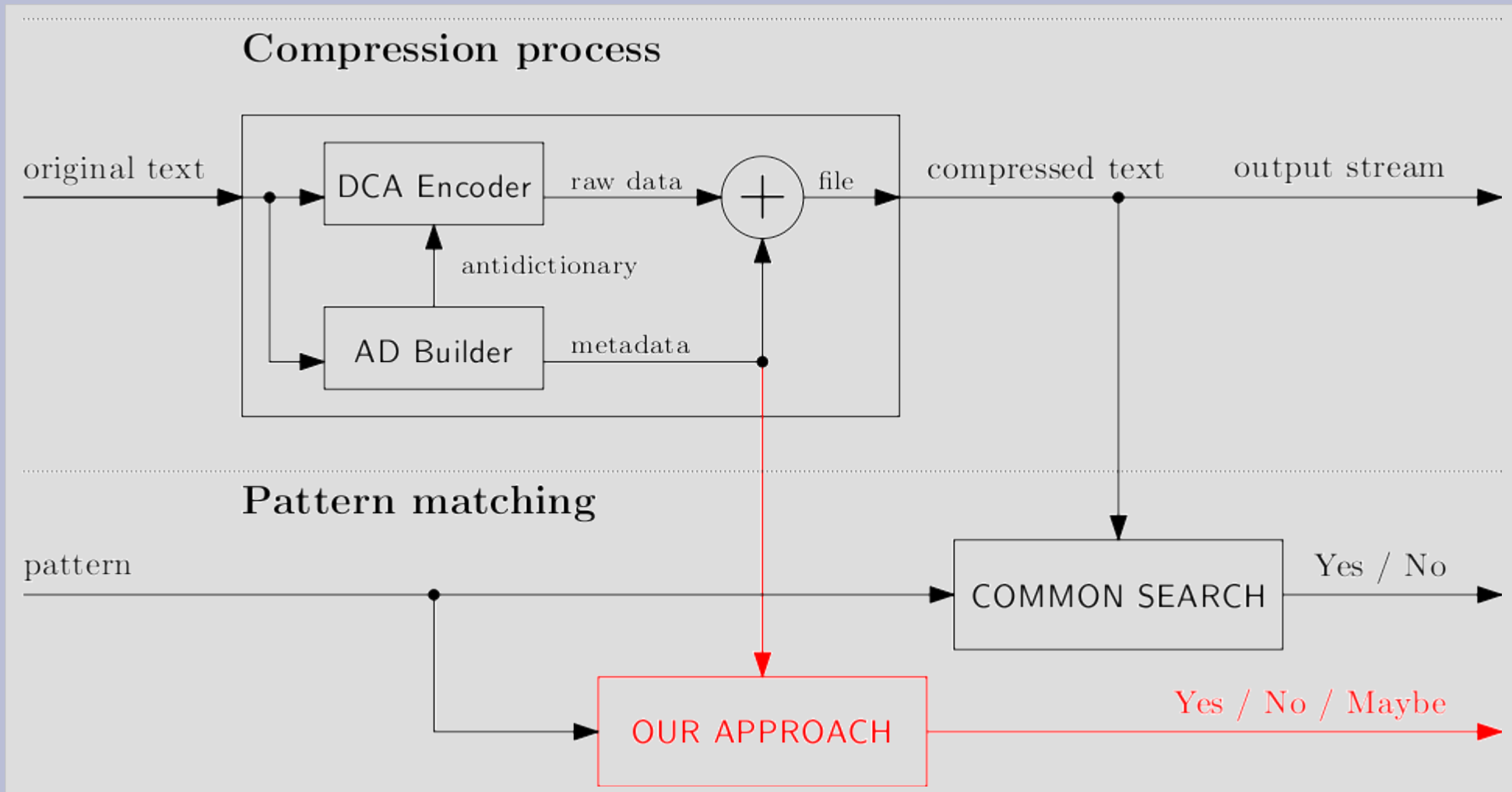
Motivace

- DCA – “Data Compression using Antiditionaries”
  - Relativně nová kompresní metoda [Crochemore, 2000]
  - Kontextová
  - Nad binární abecedou {0, 1}
  - Výpočet antislovníku = zdroj asymetrie (delší komprese)
- Možnost vyhledávání vzorků v metadatech
  - Vyhledáváme v komprimovaném textu bez dekomprese
  - Věštící metoda – odpovědi Yes / No / Maybe
- Ukázka nasazení DCA v omezeném HW
  - Málo paměti, nízký výkon CPU, pomalý přenosový kanál

Specifikace cílů

- Nastudovat metodu DCA
  - Existující implementace [Fiala 2007]
- Navrhnout a zkonstruovat vyhledávání
  - Pomocí modelů z teorie konečných automatů
- Prověřit použitelnost vyhledávání
  - Měření závislosti síly odpovědi na parametrech testů
- DCA v HW zařízení
  - Výběr platformy s ohledem na vlastnosti DCA
  - Návrh kompresního schématu a jeho implementace

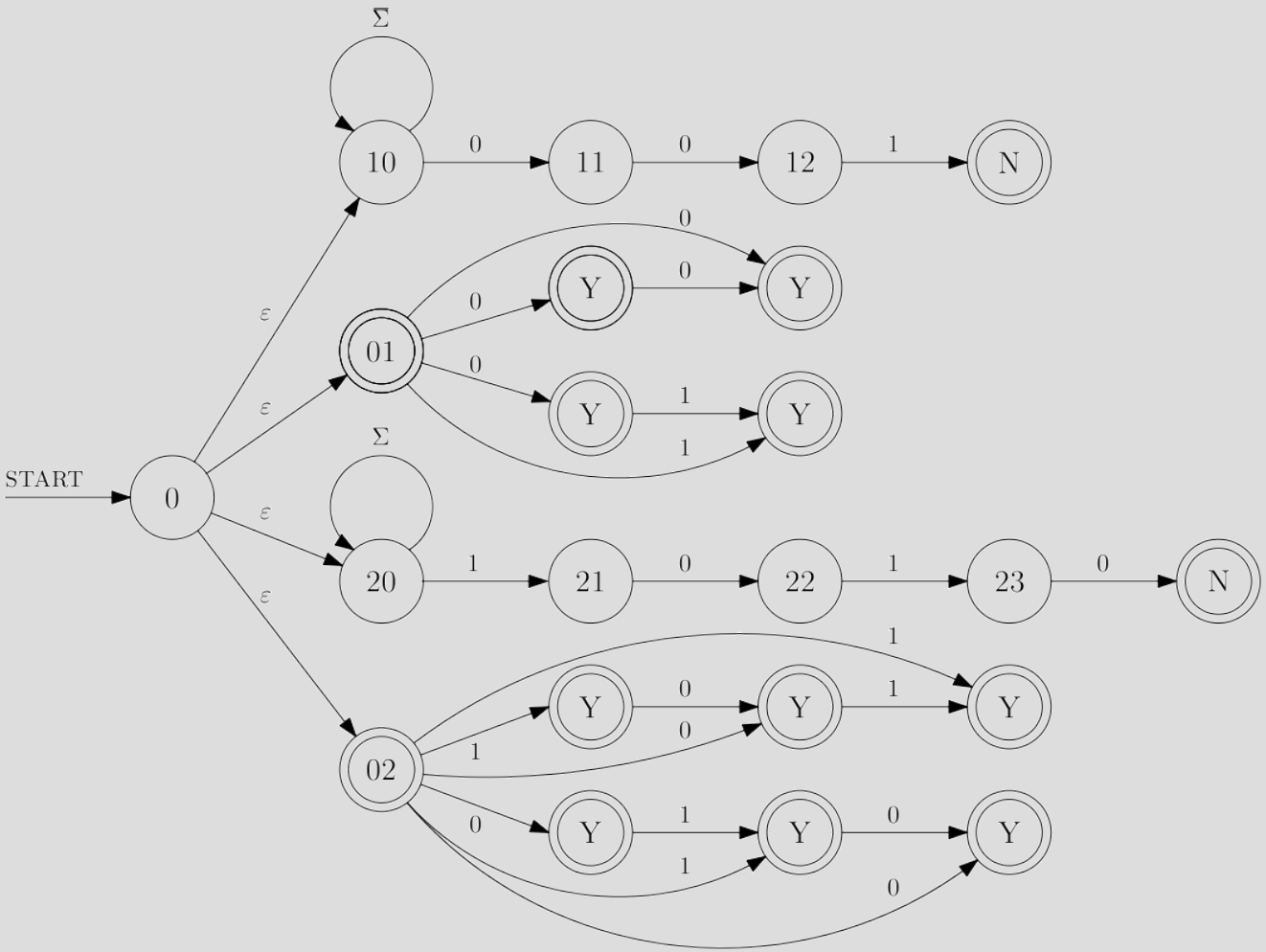
Vyhledávání v metadatech (1)



Vyhledávání v metadatech (2)

- Negativní odpověď
  - Vzorek se v původním textu nevyskytuje
  - Některé(á) antislovo je faktorem vzorku
- Pozitivní odpověď
  - Vzorek se v původním textu vyskytuje
  - Vzorek je vlastním faktorem některého(ých) antislova
- Neurčitá odpověď
  - Vzorek se v původním textu může vyskytovat
  - Ostatní případy – do antislovníku se nedostala potřebná informace pro určitou odpověď

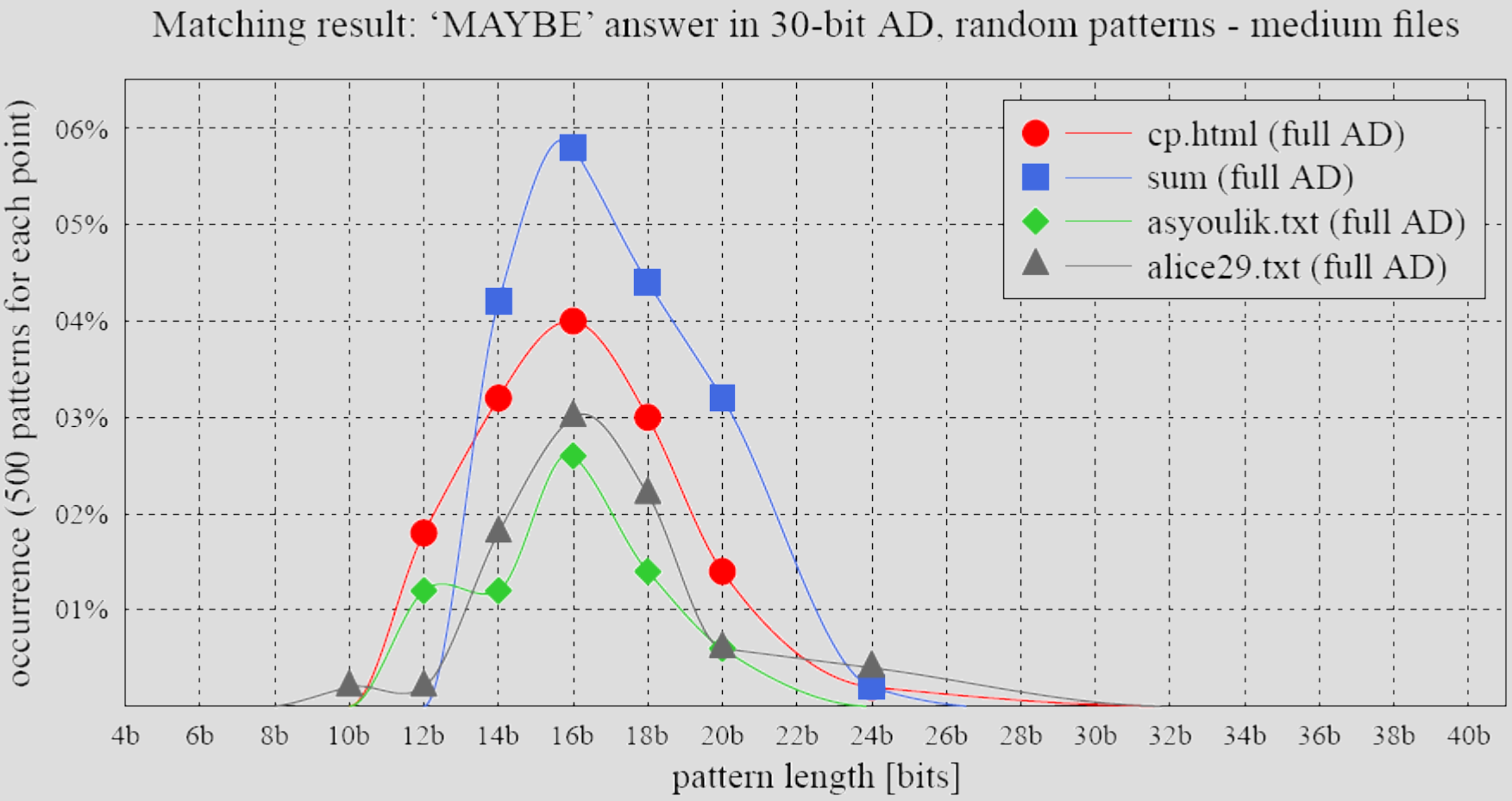
Vyhledávání – automat



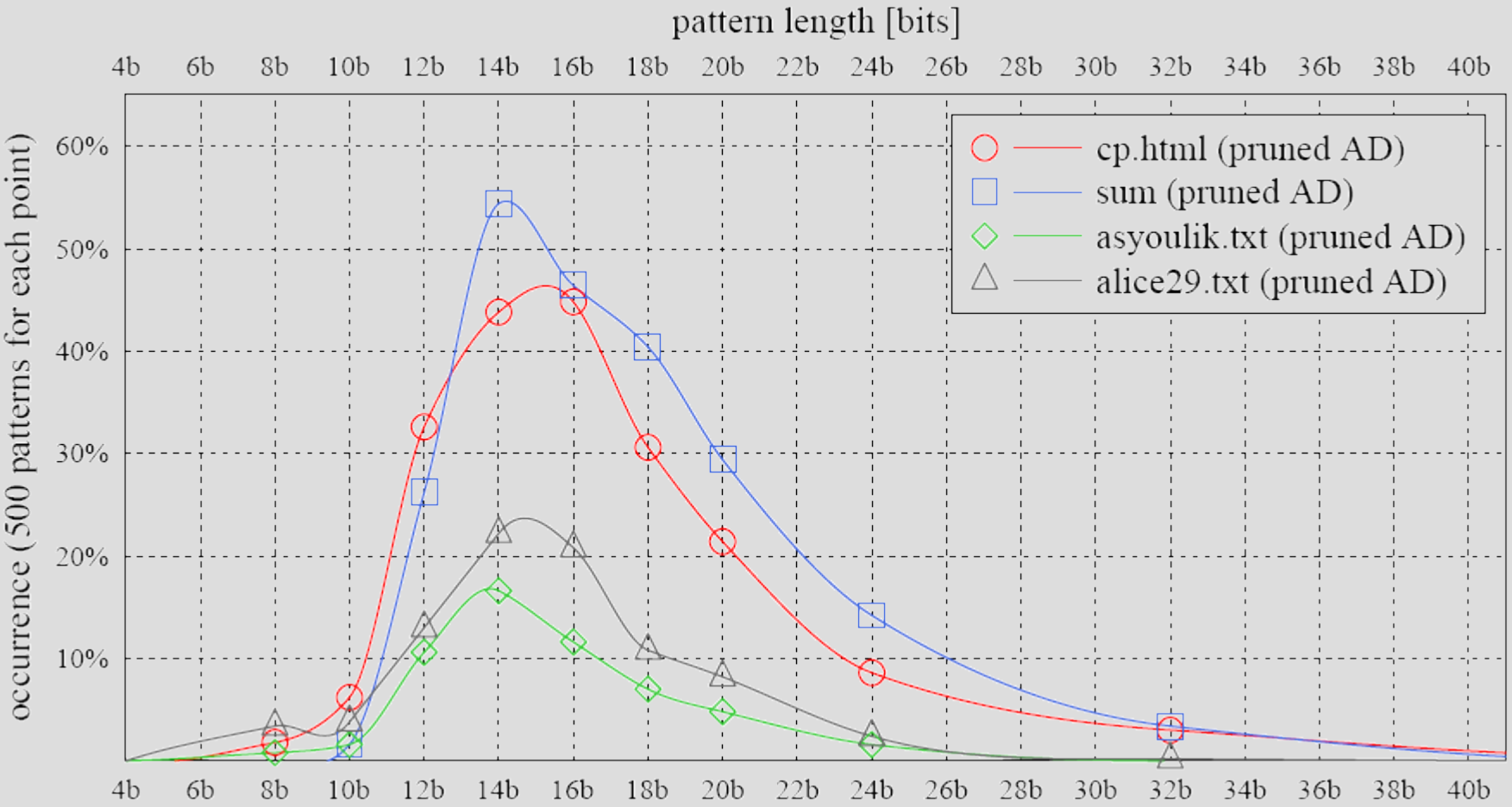
Vyhledávání – realizace

- Implementace v C++ (ISO C++ 98)
  - Simulace NKA (možnost počítání doplňujících údajů)
- Experimenty – Canterbury Corpus
  - Dimenze měření:
    - Hloubka antislovníku
    - Prořezání antislovníku
    - Velikost vstupních dat
  - Měřené závislosti (grafy):
    - Distribuce délky antislov
    - Výskyt neurčité odpovědi na náhodných/obsažených vzorcích
    - Podíl kladné a záporné odpovědi
    - Počet výsledek rozhodujících antislov pro obě určité odpovědi

Experimenty – grafy (1)



Experimenty – grafy (2)



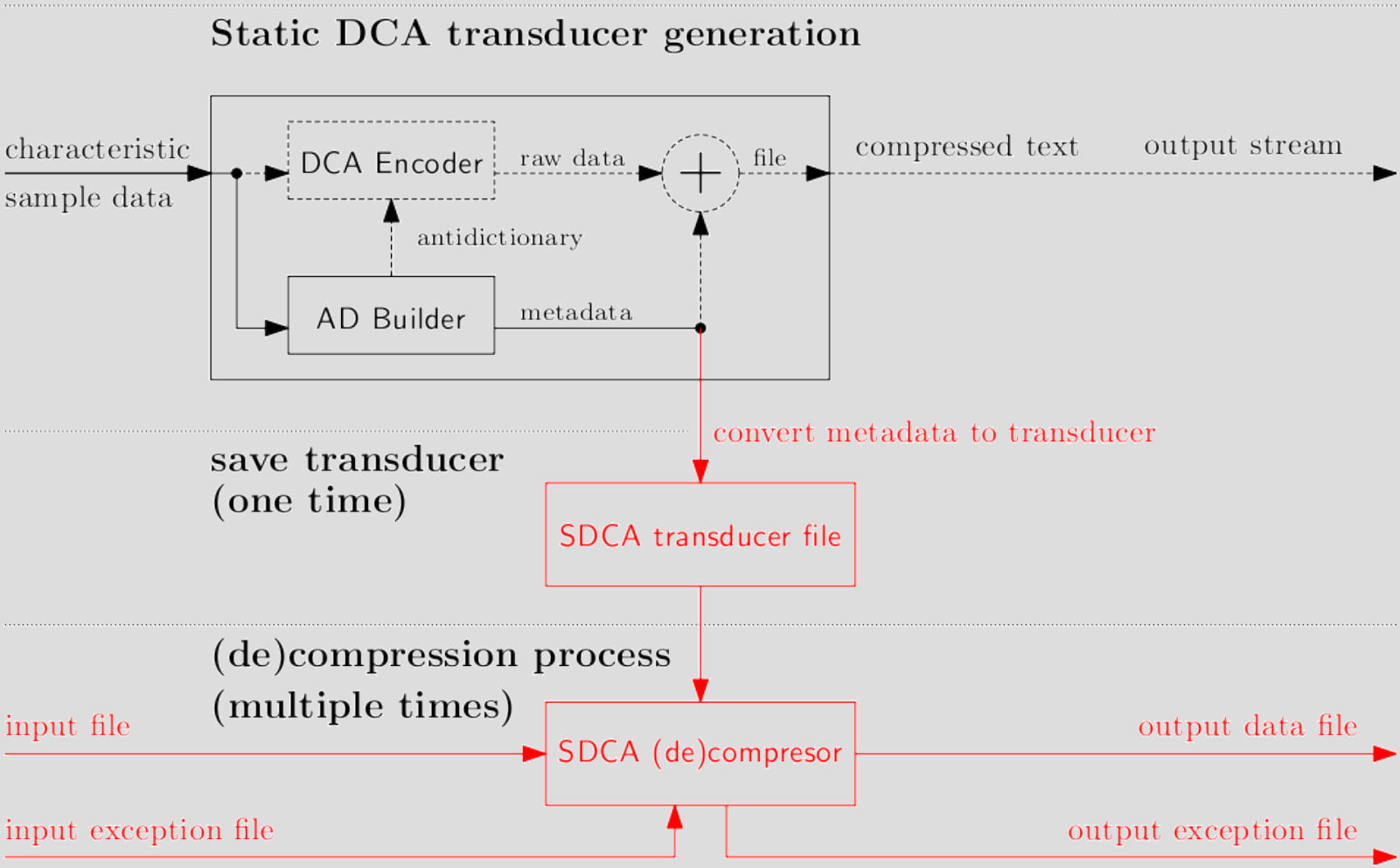
Vyhledávání – zhodnocení

- Složitost je lineární vzhledem k délce vzorku
- Vyhledávání je citlivé na prořezání AD
  - Obsáhlejší antislovník dodává více určitých odpovědí
- Vyhledávání má větší sílu při zamítání výskytu
- Vyhledávání funguje lépe pro větší soubory
- Problém zarovnání dat (do bajtů)
  - Vyhledávání (i komprese) v DCA má bitovou povahu

DCA v HW – platforma

- MJ2732VEP – Vetronics
  - Elektronická kniha jízd od firmy Princip, a.s.
  - RISC CPU STR911FAM44 (ARM), 96 kB RAM
  - Upload monitorovacích logů po GPRS (.hex)
  - Nemožnost nasadit zip kvůli omezeným prostředkům

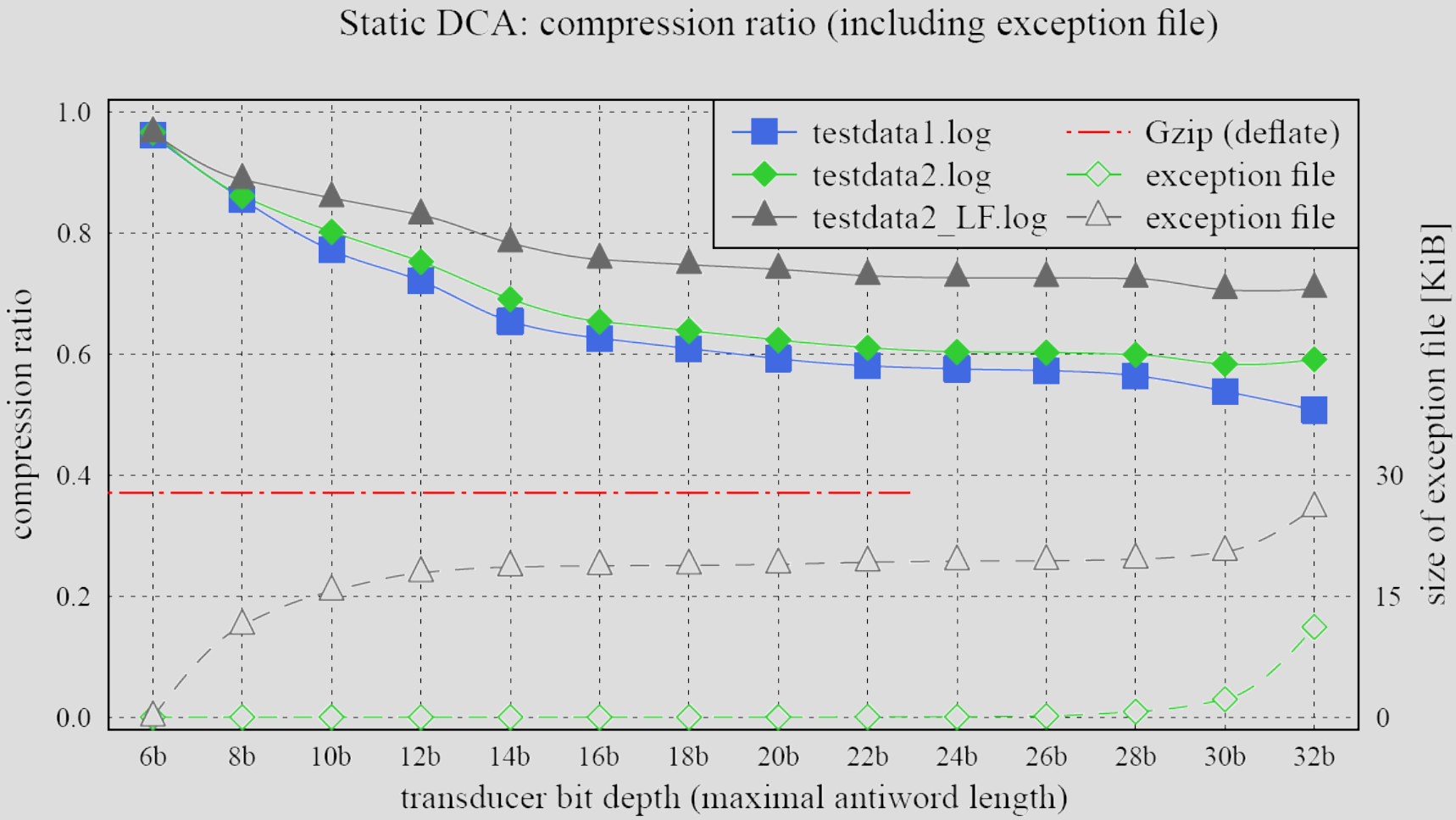
DCA v HW – schéma



DCA v HW – implementace

- Implementace v C (ISO C 99)
  - Standardizační mezivrstva pro práci se soubory
  - Multiplatformita kódu (možnost ladění v gdb na PC)
- Statické schéma
  - Výpočet antislovníku z referenčních dat (jednou na PC)
  - Generátor automatu pro statické DCA
  - Komprese/dekomprese na platformě Vetronics
  - Kódování výjimek (Fibonacciho kód)
- Volitelná spotřeba paměti
  - Bitová hloubka AD udává dosahované kompresní poměry

DCA v HW – výsledky



Závěr

- DCA v HW
  - Ve zvolené aplikaci bylo dosaženo komprese cca 0.6
  - Rychlost algoritmu řádově větší než souborový subsystém
- Splnění cílů práce
  - Seznámili jsme se s kompresní metodou DCA
  - Navržen a zkonstruován automat pro vyhledávání bitových vzorků nad metadaty (antislovníkem) DCA komprese
  - Použitelnost vyhledávání prověřena řadou experimentů
  - Nalezena HW aplikace pro ukázku DCA v praxi
  - Implementována DCA komprese se statickým schématem
  - Otestováno chování statického DCA na reálných datech

Prezentace diplomové práce

Děkuji za pozornost