AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Implementation of algorithms for fast text search and files comparison

E. Jamro, M. Wielgosz, P. Russek, M. Pietroń,
D. Żurek, M. Janiszewski, and K. Wiatr
ACC Cyfronet AGH
Dept. of Electronics AGH

Agenda

About project Synat

Hardware acceleration of big data processing

Data compression in FPGA

Ad hoc text search in FPGA

Finding text similarities

www.synat.pl

ICM Uniwersytet Warszawski/

Politechnika Wrocławska

ACK CYFRONET, Akademia Górniczo – Hutnicza

Uniwersytet Kardynała Stefana Wyszyńskiego/

Instytut Łączności, Państwowy Instytut Badawczy/

Politechnika Gdańska

Polsko Japońska Wyższa Szkoła Technik Komputerowych/

Instytut Podstaw Informatyki, Polska Akademia Nauk

PCSS IChB PAN/

Politechnika Warszawska

Biblioteka Narodowa

Uniwersytet Jagielloński

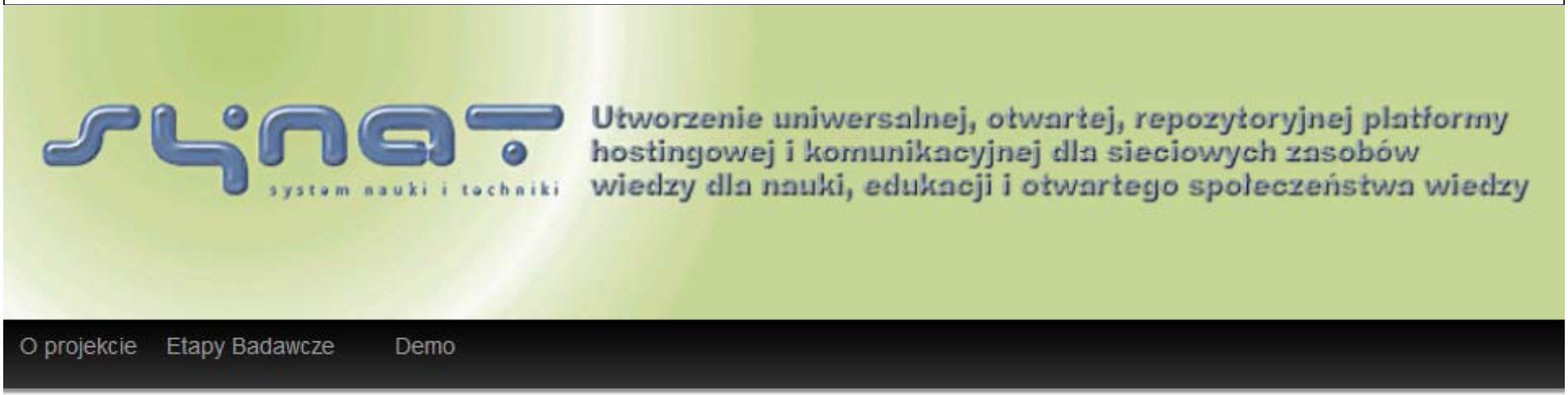
Uczelnia Łazarskiego

Naukowa i Akademicka Sieć Komputerowa

Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski

Wydział Cybernetyki, Wojskowa Akademia Techniczna

synat.cyfronet.pl



Strona Projektu SYNAT ACK Cyfronet AGH

Celem projektu jest stworzenie *uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla s ieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy*

Proponowana realizacja obejmuje szeroki zakres zadań o charakterze badawczym, podporządkowany głównemu celowi – st worzeniu kompleksowego systemu, który obejmie:

- Platformę informatyczną, realizującą całokształt funkcji użytkowych systemu,
- Podsystemy aplikacyjne, umożliwiające platformie obsługę szerokiej palety zasobów treściowych, z zapewnieniem wysokie go poziomu skalowalności, a także interoperacyjności w układzie międzynarodowym,
- Podsystemy generyczne umożliwiające integrację nowych klas przyszłych aplikacji,
- Podsystem nowych modeli komunikowania naukowego i otwartych społeczności wiedzy, obejmujący również program upo wszechniania i promocji adresowany do całego społeczeństwa,
- Zbiór propozycji modeli prawnych umożliwiających rozwój nowych otwartych modeli komunikowania w nauce, edukacji i obs zarze dziedzictwa kulturowego,
- Model operacyjny, zapewniający trwałość systemu, a także podejmujący kwestie możliwych obszarów jego komercjalizacji.

Strona projektu: www.synat.pl

Deflate
Text compression in FPGA

Standard employed in .zip / .gz

Combination of two methods:

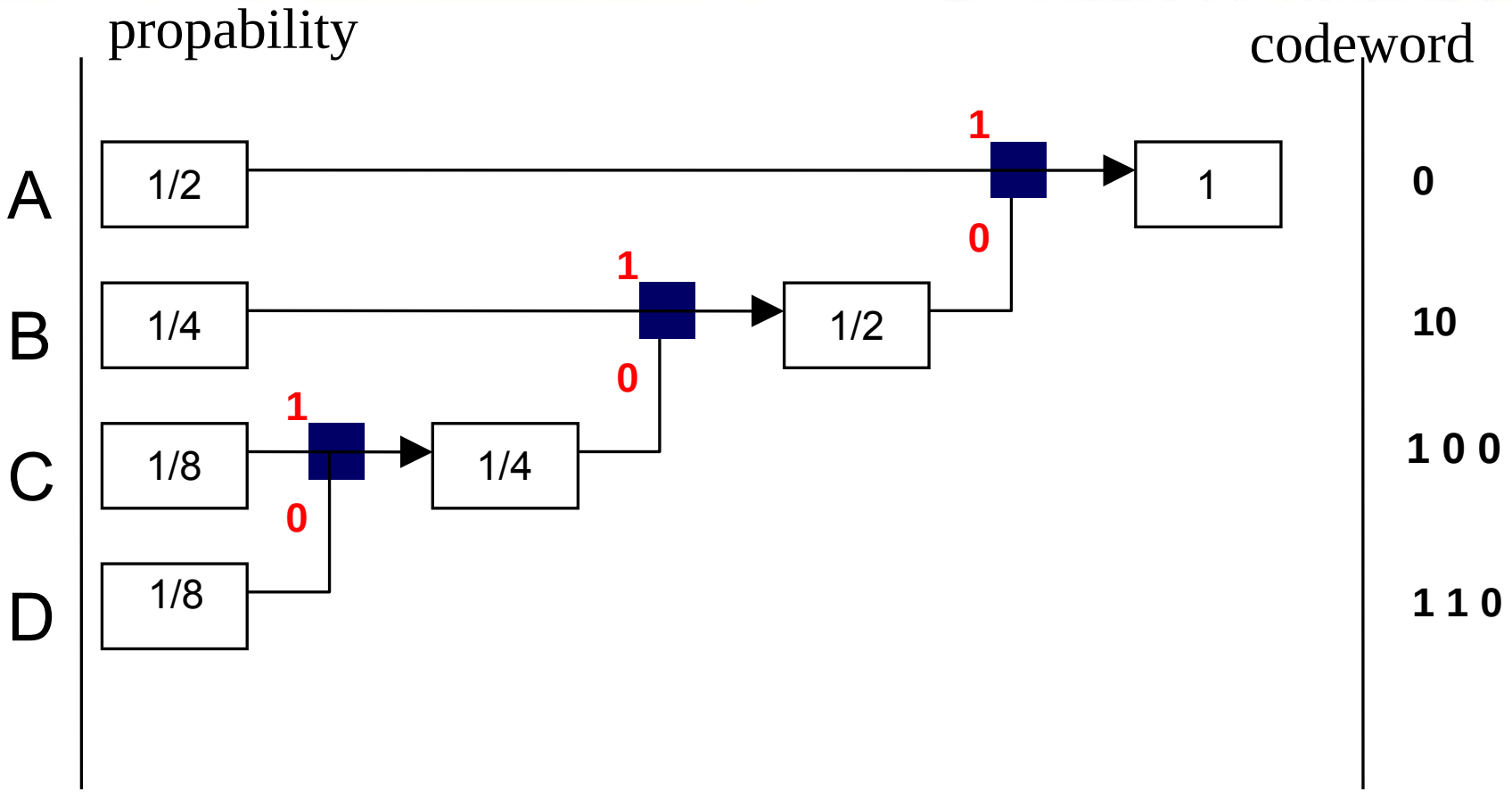
- LZ77 / LZSS dictionary

  ABCDABCE ➔ ABCD(length=3, distance=4)E

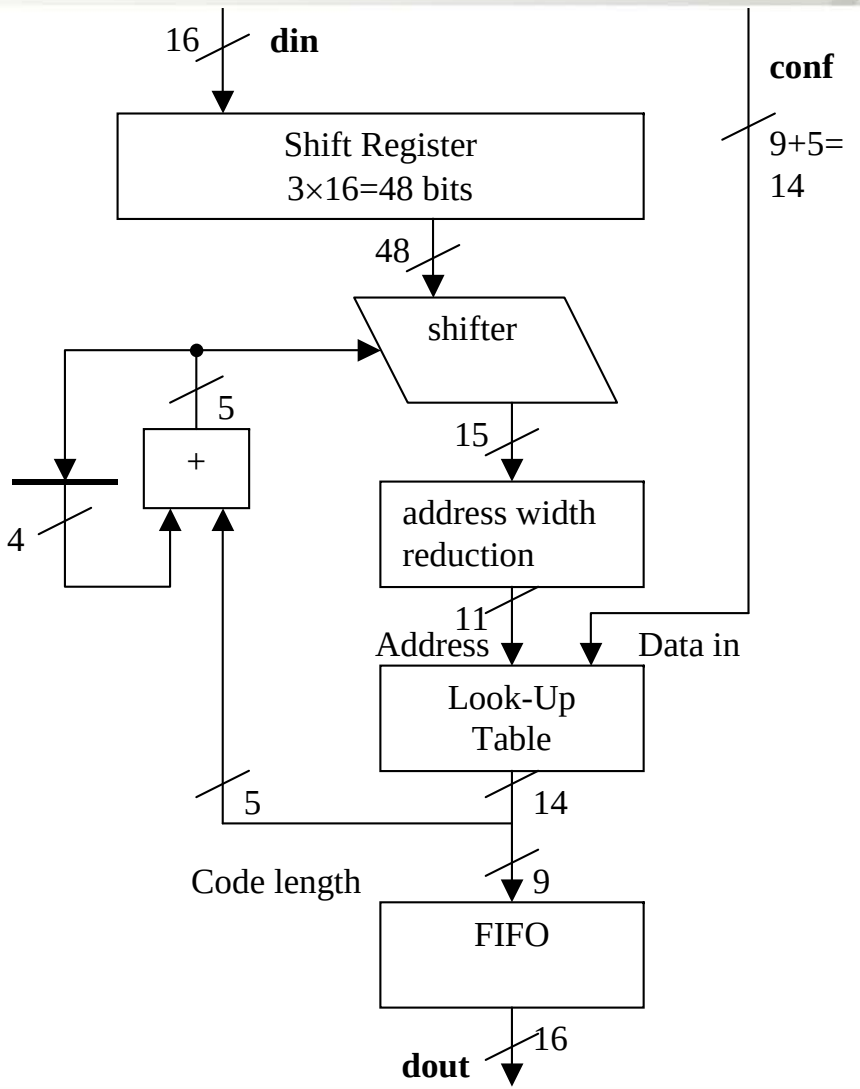
  ABABABABA ➔ AB(length=7, distance=2) - RLE

- Huffman Coding (entropy based coding)

Huffman coding



Huffman decoding



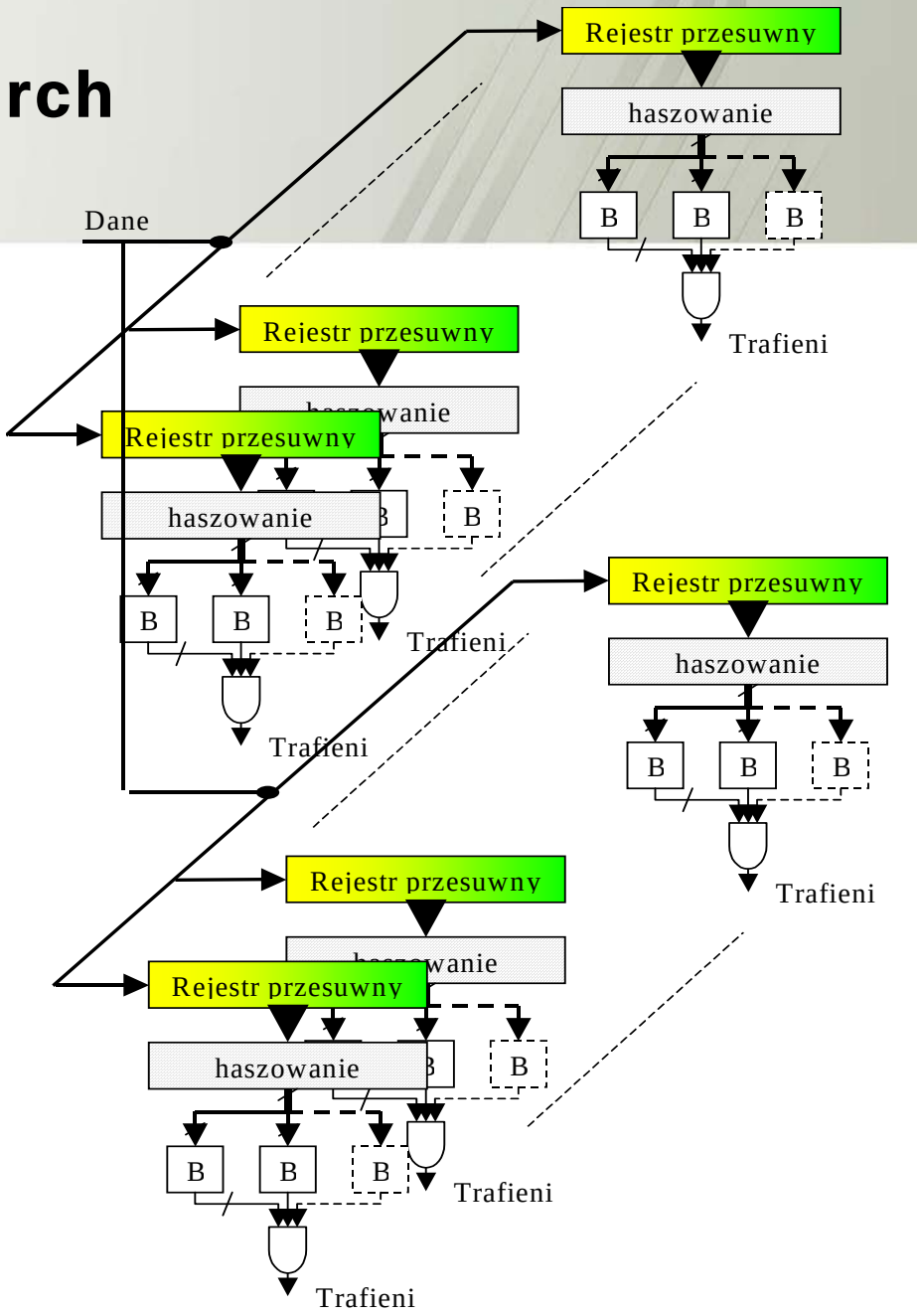
Big data ad hoc search

Bloom Filter

Throughput: 1.6GB/s – limited by the NUMAlink

Strongly Parallel 100 000 patterns simultanously

Speed-up 200 in comparison to Itanium2 1,5GHz and grep – number of patterns=50



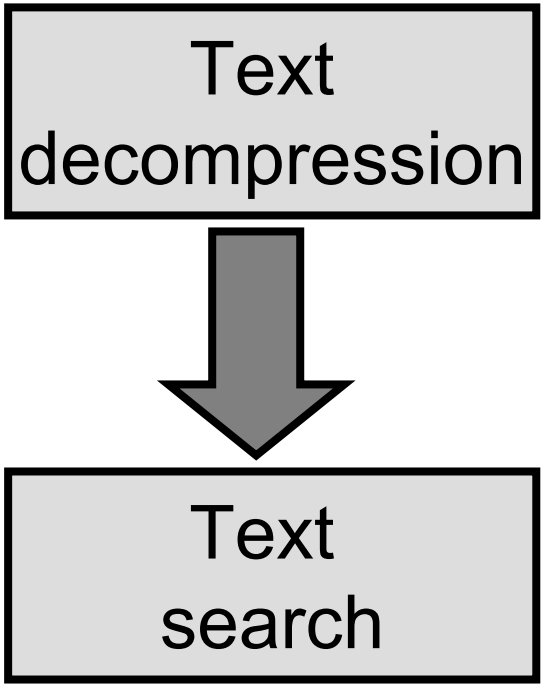
Combination of decompression and text search

FPGA implemented decompression practically not limits the throughput, but many independent streams must be used!!

Text compression ratio: 2-4 times

Summing up:

- Less HDD / memory space is required
- Much more data can be delievered to FPGA (after decompression)



Practical approach to big text search

Index search

implemented in software

Ad hoc search

implemented in FPGA

Plagiarism Detection
Text similarity search

Plagiarism Detection

www.cyfronet.pl/synat