

# Hardware implementation LZ4

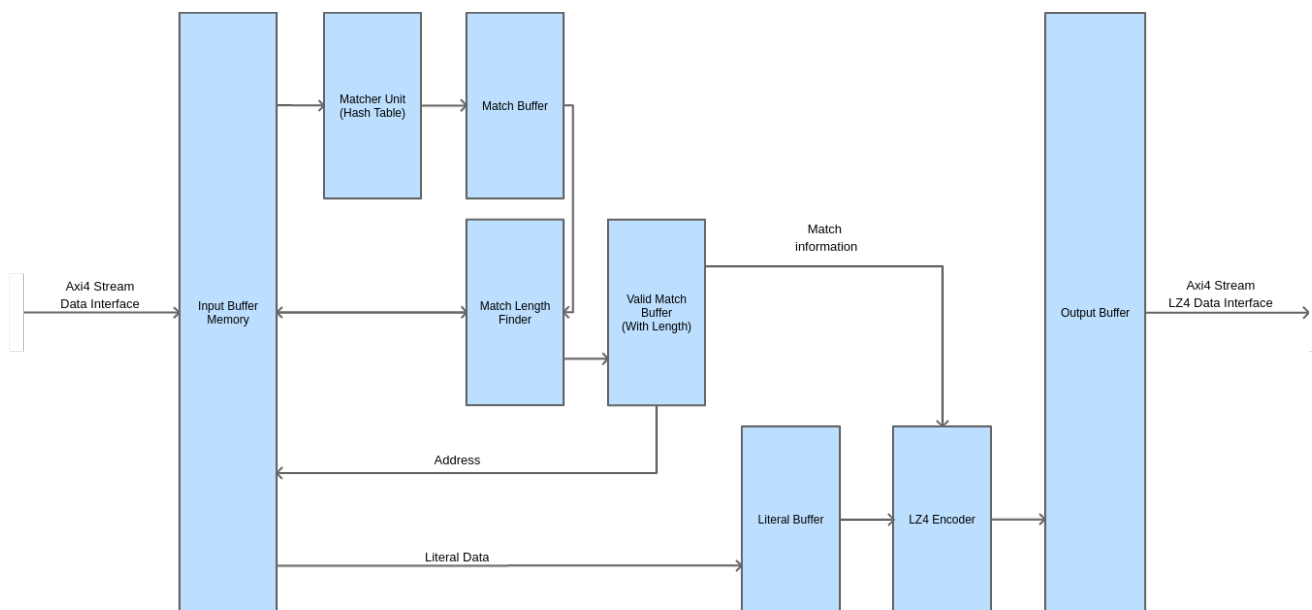
Tomáš Beneš

30. října 2018

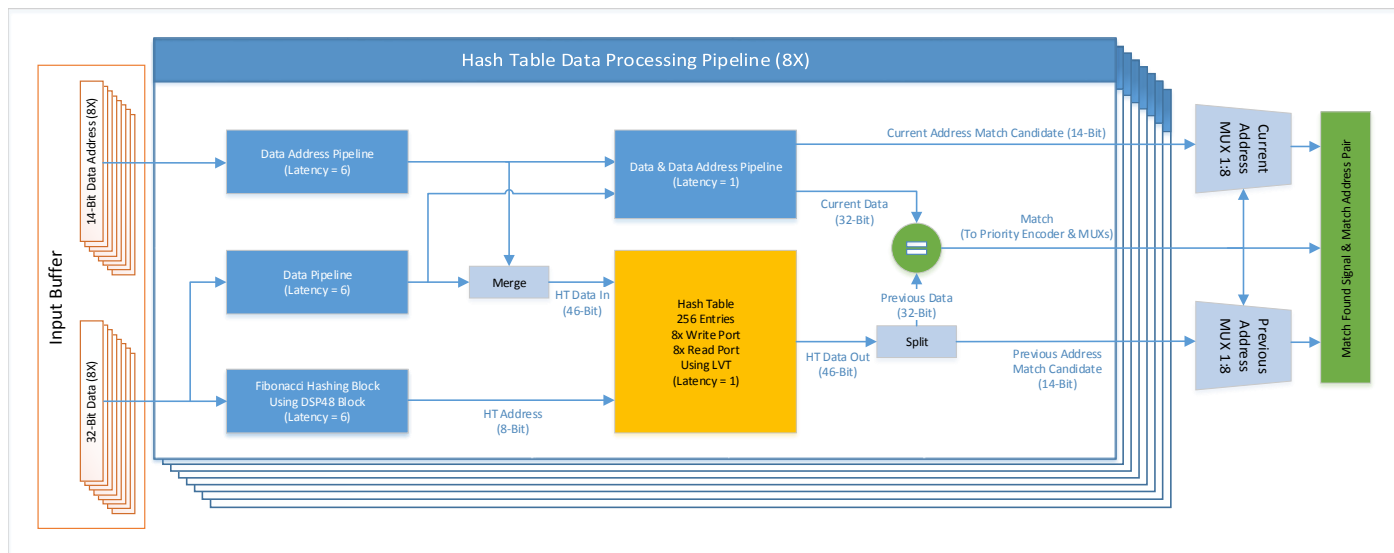
# Hardware hierarchy

Práce navazuje na výlet z roku 2017, kde byla navržena a otestována jednotka pro hledání schod pro LZ4, tato jednotka byla použita a vylepšena při integraci do kompresního bloku. Jednotka je rozdělena do 3 částí, kde se u každé jednotky zkoumala její propustnost v nejhorsím případě, aby byla splněna podmínka pro 10Gb/s propustnost.

Každá část pracuje nezávisle na sobě a má výstupní buffer pro výsledky, aby se udržel koncept pipeline. Při kterém má jednotka propustnost jako je nejnižší propustnost řetězených jednotek.



Obrázek 1: Blokové schéma kompresního bloku



Obrázek 2: Blokové schéma jednotky hledání schod

## Hashtable Matcher compression ratio

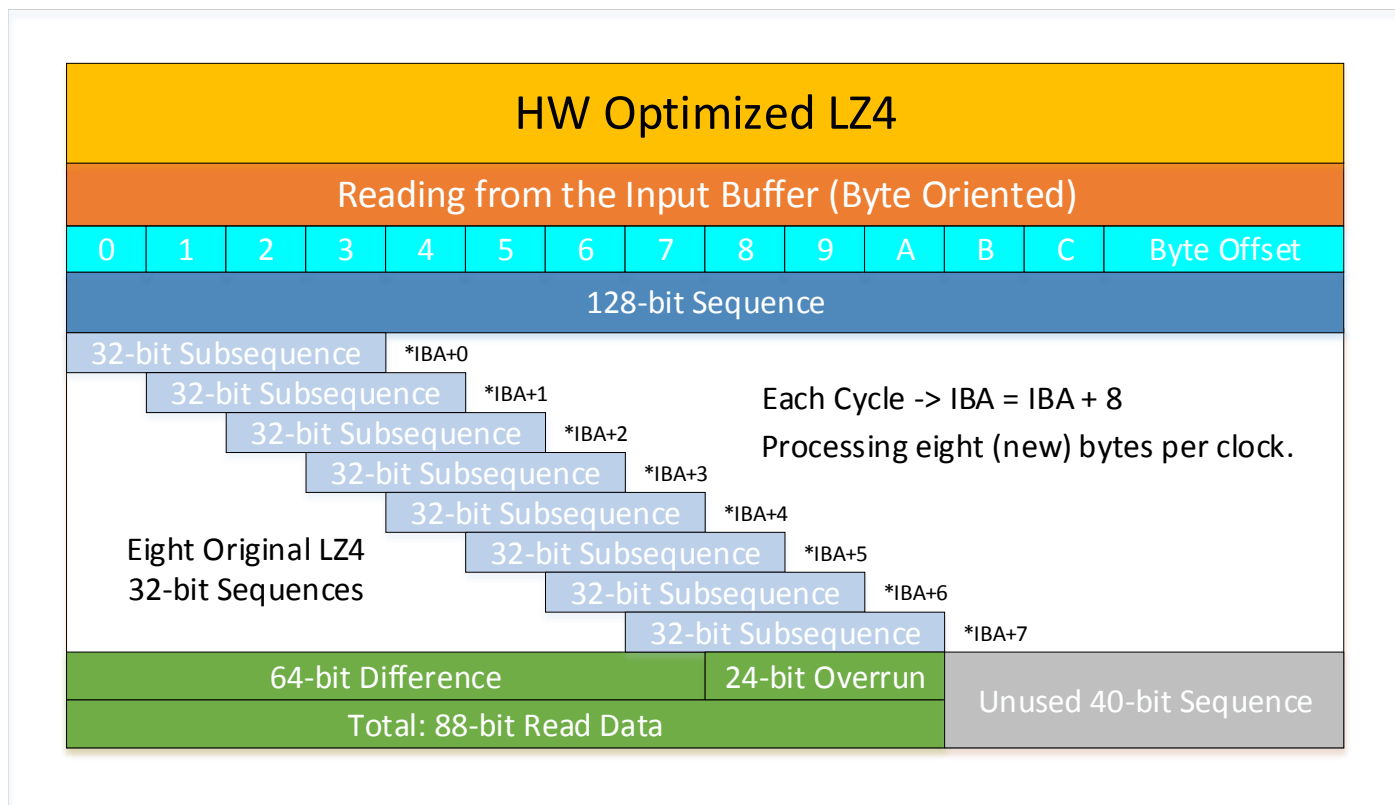
Pro publikaci článků v oblasti komprese se ukázalo, že článek musí obsahovat alespoň nějakou studii o kompresním poměru. Článek o jednotce hledání schod byl doplněn o experimentální studii kompresního poměru při použití hashovací tabulky o dané velikosti.

Tato studie byla implementovaná jako softwarový funkční model kompresního algoritmu lz4 za pomoci hash-tabulky a následující zpracování schod greedy způsobem, to jest u schod, které následují přímo po sobě se bere nejdelší. Jednotka hledání schod se nezastavuje, takže u schod delší jak 4 byty zaznamenává po sobě jdoucí schody.

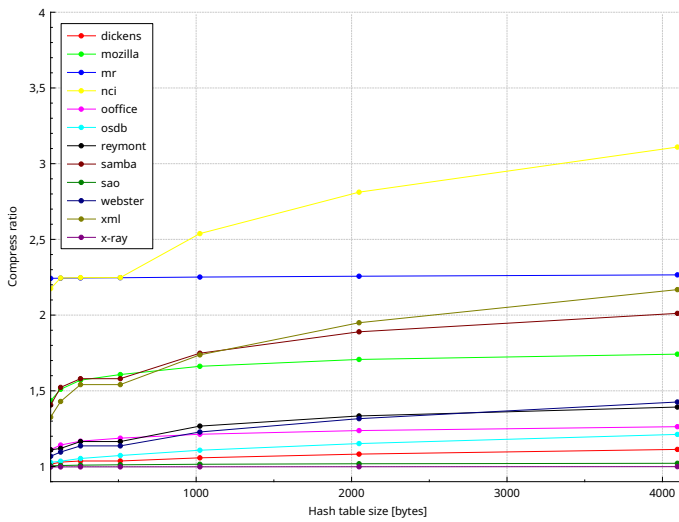
Vstupem pro experimenty jsou známé corpusy Canterbury, Calgary a Silesia corpus. Na grafech je vidět závislost kompresního poměru na velikosti hash-tabulky, kde velikost tabulky je pro naši jednotku hledání schod lineárně závislá

na množství zdrojů v hardware (využití BRAM jednotek). To nám dává volnost v nastavení kompresního poměru podle dostupných zdrojů v FPGA/ASIC. V tabulkách je porovnání s oficiální softwarovou implementací lz4, která používá více pokročilé techniky hledání schod.

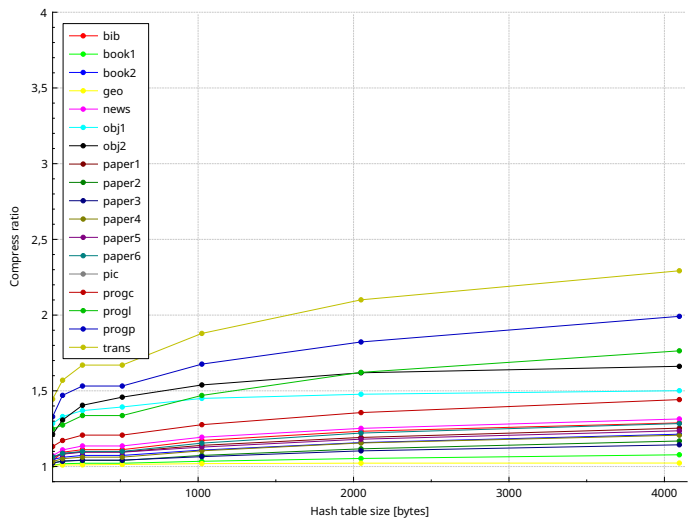
Výsledek experimentálně ověřené poměry se na jednoduchost implementace blíží částečně k oficiálnímu řešení a pro implementaci v hardware s propustností 10Gb/s jsou více než dostačující.



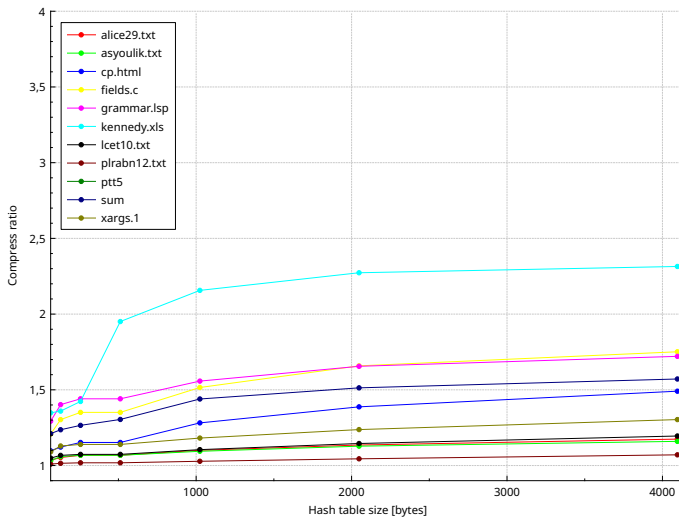
Obrázek 3: Způsob zpracování schod



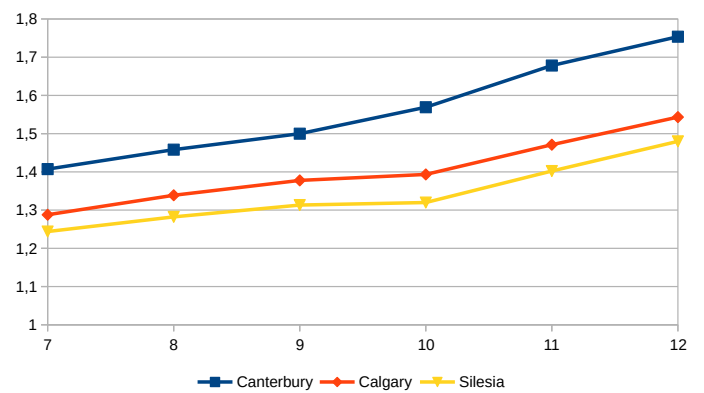
(a)



(b)



(c)



(d)

Obrázek 4: Kompresní poměr naměřen pro corpus(a - Silesia, b - Calgary, c - Canterbury, d - Average)

File	64	128	256	512	1024	2048	4096	Reference
dickens	1.02	1.03	1.04	1.04	1.06	1.08	1.11	2.58
mozilla	1.43	1.51	1.57	1.61	1.66	1.71	1.74	1.93
mr	1.24	1.24	1.25	1.25	1.25	1.26	1.27	1.83
nci	2.18	2.25	2.25	2.25	2.54	2.81	3.11	6.05
ooffice	1.11	1.14	1.17	1.19	1.21	1.24	1.26	1.41
osdb	1.03	1.04	1.05	1.07	1.11	1.15	1.21	1.9
reymont	1.11	1.12	1.16	1.16	1.27	1.33	1.39	2.08
samba	1.41	1.52	1.58	1.58	1.75	1.89	2.01	2.79
sao	1	1.01	1.01	1.01	1.01	1.02	1.02	1.06
webster	1.07	1.1	1.14	1.14	1.23	1.32	1.43	2.05
xml	1.33	1.43	1.54	1.54	1.74	1.95	2.17	4.35
x-ray	0.997	0.997	0.998	0.998	0.998	0.998	0.999	1.00

Tabulka 1: Kompresní poměr Silesia

File	64	128	256	512	1024	2048	4096	Reference
alice29.txt	1.04	1.05	1.07	1.07	1.1	1.14	1.17	1.71
asyoulik.txt	1.03	1.06	1.07	1.07	1.09	1.13	1.16	1.57
cp.html	1.1	1.12	1.15	1.15	1.28	1.39	1.49	2.06
fields.c	1.21	1.3	1.35	1.35	1.52	1.66	1.75	2.13
grammar.lsp	1.29	1.4	1.44	1.44	1.56	1.66	1.72	1.92
kennedy.xls	1.35	1.36	1.42	1.95	2.16	2.27	2.31	2.74
lcet10.txt	1.05	1.07	1.07	1.07	1.1	1.15	1.19	1.82
plravn12.txt	1.01	1.01	1.02	1.02	1.03	1.04	1.07	1.47
ptt5	4.1	4.3	4.5	4.7	5	5.1	5.2	5.9
sum	1.21	1.24	1.27	1.3	1.44	1.51	1.57	2.03
xargs.1	1.09	1.13	1.14	1.14	1.18	1.24	1.3	1.57

Tabulka 2: Kompresní poměr Canterbury

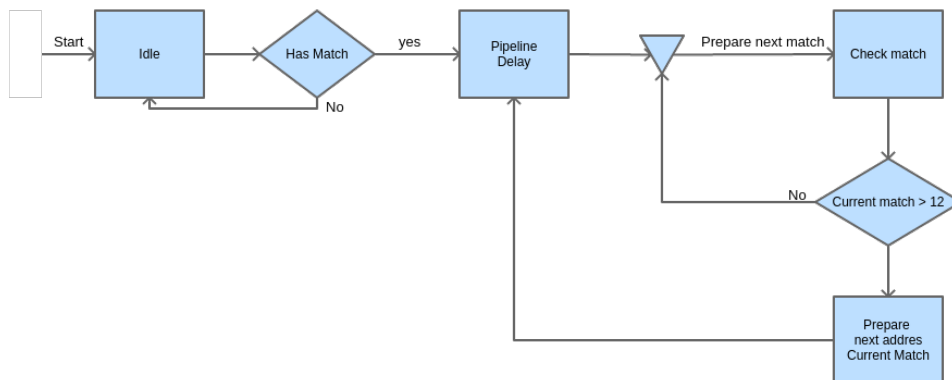
File	64	128	256	512	1024	2048	4096	Reference
bib	1.04	1.09	1.11	1.11	1.17	1.23	1.29	1.96
book1	1.01	1.02	1.02	1.02	1.03	1.05	1.08	1.47
book2	1.05	1.06	1.07	1.07	1.11	1.16	1.21	1.83
geo	1.01	1.01	1.01	1.01	1.02	1.02	1.02	1.04
news	1.07	1.11	1.14	1.14	1.19	1.25	1.31	1.69
obj1	1.28	1.33	1.37	1.39	1.45	1.48	1.5	1.66
obj2	1.21	1.31	1.4	1.46	1.54	1.62	1.66	2.09
paper1	1.06	1.09	1.1	1.1	1.14	1.19	1.25	1.83
paper3	1.02	1.04	1.04	1.04	1.07	1.1	1.14	1.64
paper4	1.03	1.05	1.06	1.06	1.1	1.16	1.21	1.56
paper5	1.06	1.08	1.09	1.09	1.13	1.18	1.24	1.60
paper6	1.06	1.09	1.1	1.1	1.15	1.22	1.28	1.84
pic	4.1	4.3	4.5	4.7	5	5.1	5.2	5.9
progc	1.13	1.17	1.21	1.21	1.28	1.36	1.44	1.89
progl	1.25	1.27	1.34	1.34	1.47	1.62	1.76	2.65
progp	1.33	1.47	1.53	1.53	1.68	1.82	1.99	2.63
trans	1.45	1.57	1.67	1.67	1.88	2.1	2.29	3.11

Tabulka 3: Kompresní poměr Calgary

## Match length finder

Výstup z jednotky pro hledání schod se musí projít v paměti pro získání informace o délce schody, zároveň jednotka musí reagovat na po sobě jdoucí schody, aby dosáhla propustnosti 10Gb/s. V nejhorším případě jednotka musí zpracovávat schody o délce 4 byty s rozestupem jednoho bytu, z toho plyne, že musí zpracovávat při naší cílené frekvenci 1 schodu o délce 4(8 bytu celkem) za jeden takt.

Tohoto bylo docíleno pomocí předpokladu, že schody délky 4-12 jsou odbaveny za jeden hodinový takt, jednotka vystavuje adresu další schody v poradí ještě před ověřením současné schody, protože paměť reaguje se zpožděním jednoho taktu. Pokud je shoda delší musí se pipeline do paměti vyprázdnit toto trvá 2 takty v tomto případě se dostáváme na shodu délky 13-21 bytů kde je potom propustnost akorát 10Gb/s, protože tím odbavíme 26-42 bytů za 3 hodinové takty(8.6 bytu za takt).



Obrázek 5: Diagram chování jednotky hledání schod

## Práce

Výstupní automat pro kódování formátu je v současné době ve vývoji, a v době má některá omezení v propustnosti. V budoucím výzkumu bude tento automat předělán, aby se skládal z několika jednotek, aby dokázal kódovat formát lz4. Dále probíhá práce na chodu jednotky v plynulé pipeline režimu, v současné době dokáže jednotka zakódovat block dat v režimu paketovém (spustí se komprese až při přenesení dat do jednotky). Současné výsledky práce jsou kompletní funkční simulace jednotky v paketovém režimu, kde byla dosažena propustnost **7.1Gb/s** V budoucí práci je v plánu předělat kódovací jednotku aby jsme dosáhli plné rychlosti 10Gb/s.