Reconfigurable Convolutional Kernels for Neural Networks on FPGAs

Martin Hardieck, Martin Kumm, Konrad Möller, Peter Zipf

University of Kassel

27th ACM/SIGDA International Symposium on Field-Programmable Gate Arrays

Monterey, California

24.02.2019

Motivation

Some Inputs

- Convolutional layers are mostly weight stationary
- Many upcoming NN-accelerators
- Designs mostly stop at arithmetic level and use DSPs
- Our background is low level arithmetic optimisation

Goal

- Create plugin Convolutional Core for NN-accelerators
- Take usage of mostly static weights
- Save resources

Why not use Embedded Multipliers?

- Limited in quantity
- Scale bad with low wordsizes
- Should be saved for dense layers
- Or other tasks ...

Proposed Architecture

Idea

- Use Constant Multiplictaion
  - Resource reduction
- Make it reconfigurable
  - Adapt for new weights
  - Reconfiguration time
- Code generator
  - Easy adaption for any case
  - Free available: FloPoCo [7, 9]

Proposed Architecture

Nothing New?

- All concepts are kown
- Never combined in this way before

Used Concepts

- Generic LUT-based constant multiplication (KCM) [4, 5]
- Compressor trees [25, 26]
- Run-time reconfigurable LUTs (CFGLUT)
- Online computation of LUT content [22]
- Faithful rounding [8]
- Shadow LUTs

Generic LUT-based Constant Multiplication [4, 5]

LUT-based KCM Approach

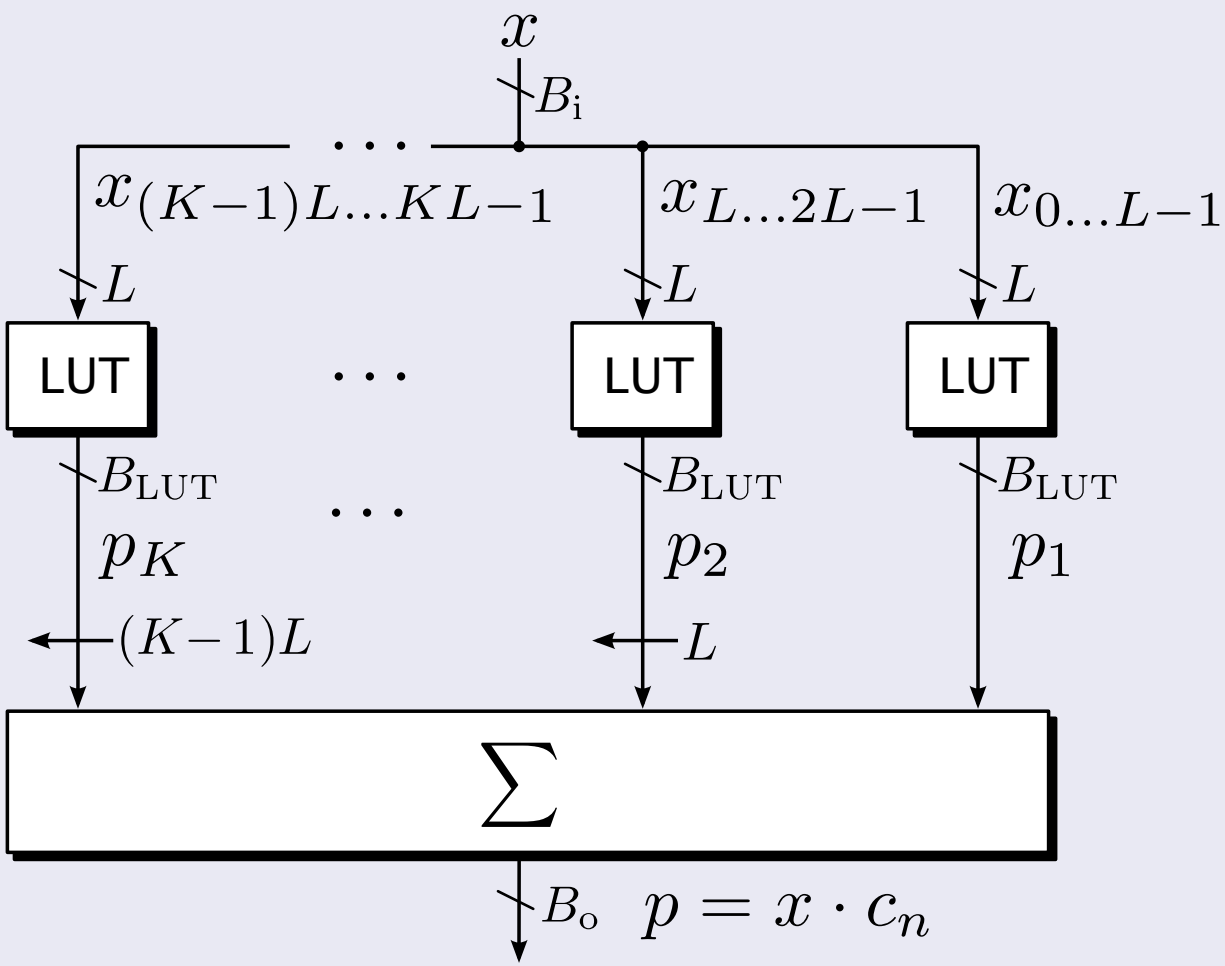
- Ken Chapman
- Represent multiplication by partial products
- Choose size of products to fit in LUT
- Summation of all bit shifted partial products (SOP)

Final Summation

- Compressor tree instead of adder tree
- Same compressor tree for all coefficients

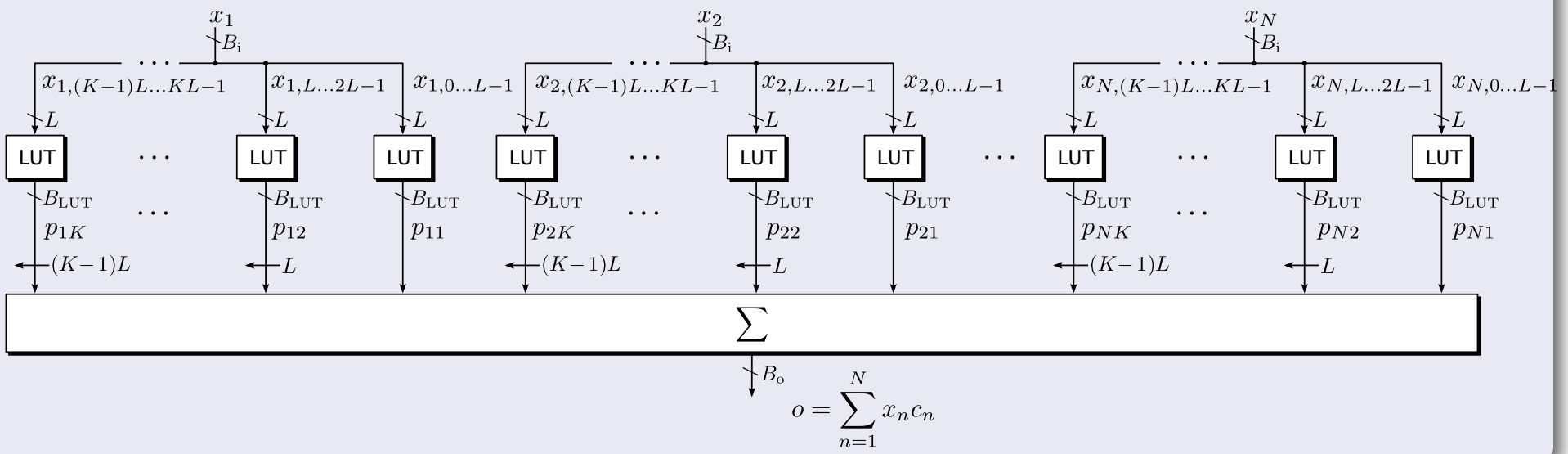
Generic LUT-based Constant Multiplication [4, 5]

SOP KCM Example



Generic LUT-based Constant Multiplication [4, 5]

SOP KCM Example



Resulting Wordsize Truncation

SOP Result

- Much bigger then input
- Truncation to output wordsize needed

Truncation to output wordsize

- Classic truncation needs complete result
- Faithful rounding [8]
  - Same accuracy as truncation
  - Not necessarily same output
  - Use just view guard bits

Low Wordsize vs. Reconfiguration Data

Coefficent Wordsize

- Often around 8 bit
- Loading coefficients is one primary bottlenek

SOP KCM Reconfiguration Data

- Reconfiguration data for MSB LUT (min 32 bit)
- Reconfiguration data for all other LUTs (min 32 bit)
- $\rightarrow$ Configuration data for each coefficient (min 64 bit)

Low Wordsize vs. Reconfiguration Data

Coefficent Wordsize

- Often around 8 bit
- Loading coefficients is one primary bottlenek

SOP KCM Reconfiguration Data

- Reconfiguration data for MSB LUT (min 32 bit)
- Reconfiguration data for all other LUTs (min 32 bit)
- $\rightarrow$ Configuration data for each coefficient (min 64 bit)

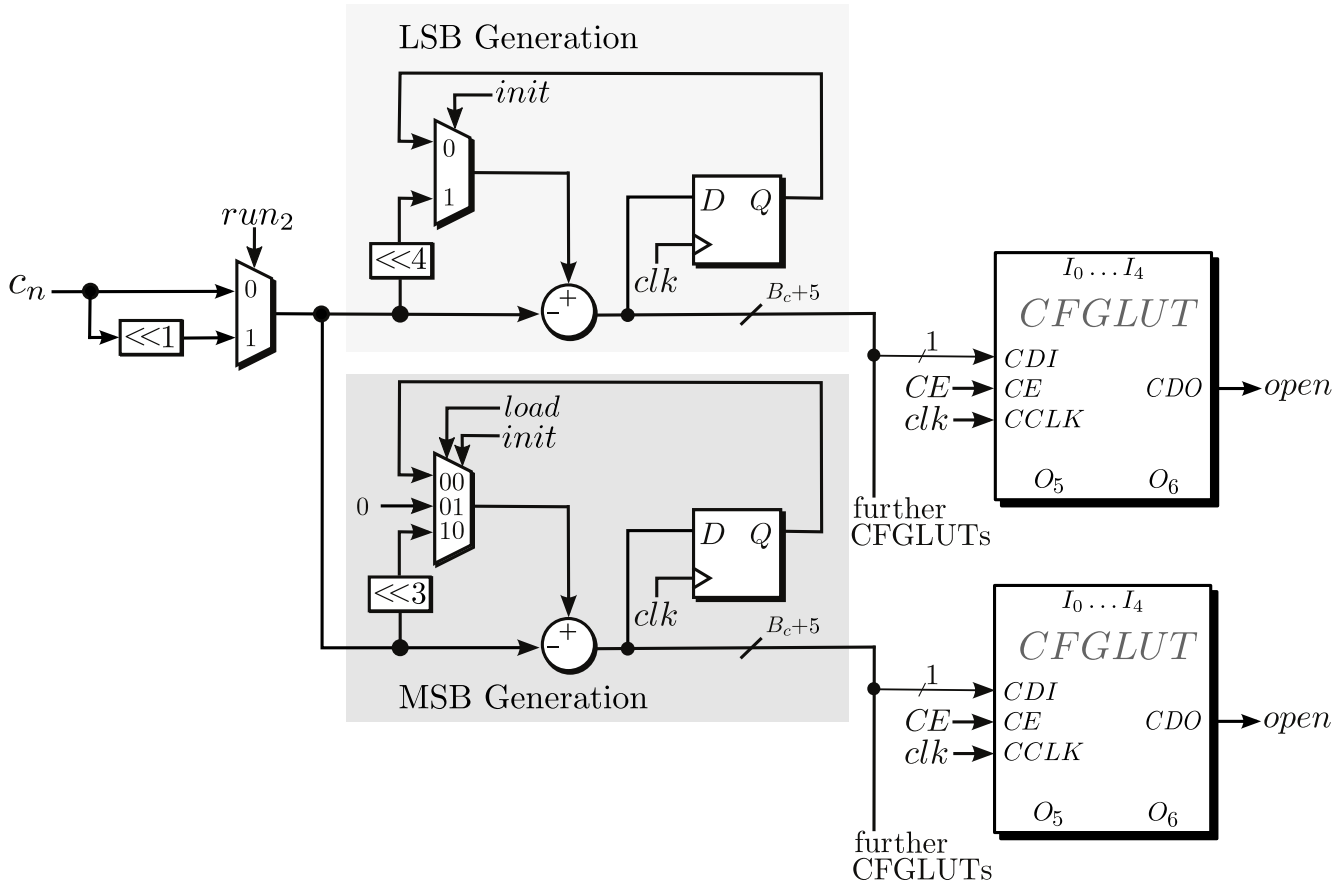
Solution

- Compute the reconfiguration data online

Online Computation of Reconfiguration Data

Configuration Circuit

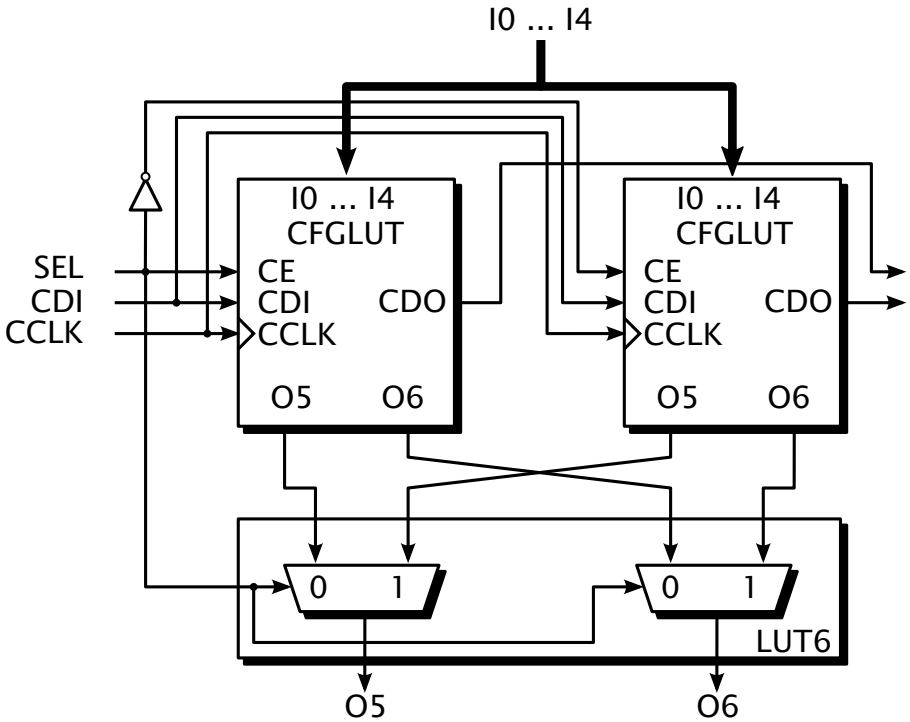
- Generate reconfiguration data dependend on the coefficent
- No change in memory requirements



How to Hide Reconfiguration Time?

Shadow LUTs

- Configure unused LUTs (in 32 cicles)
- Switch without delay (in 0 cicles)
- $\rightarrow$ Zero configuration time overhead
- Does not double resources

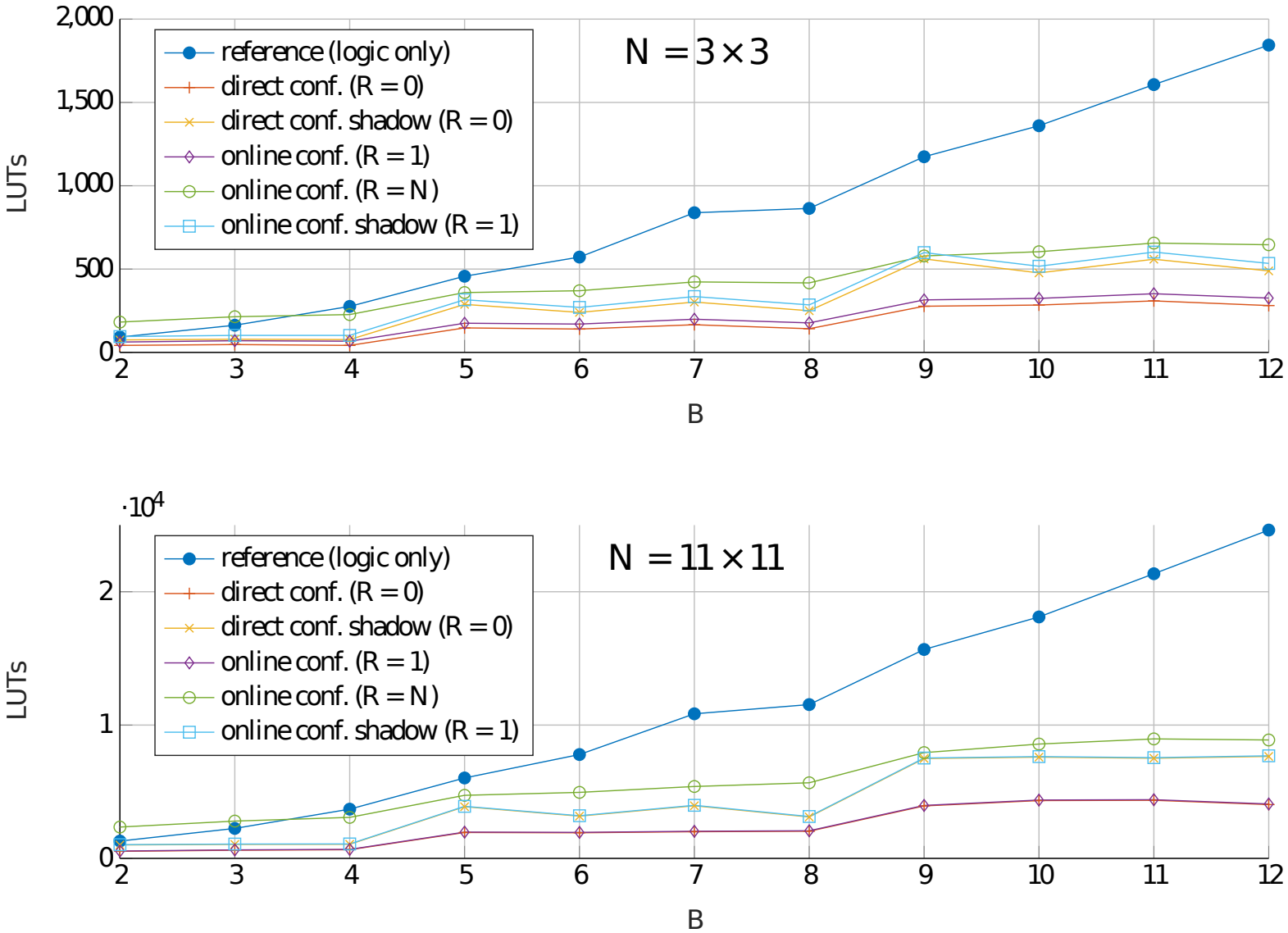


Results: Synthesis of Proposed Design

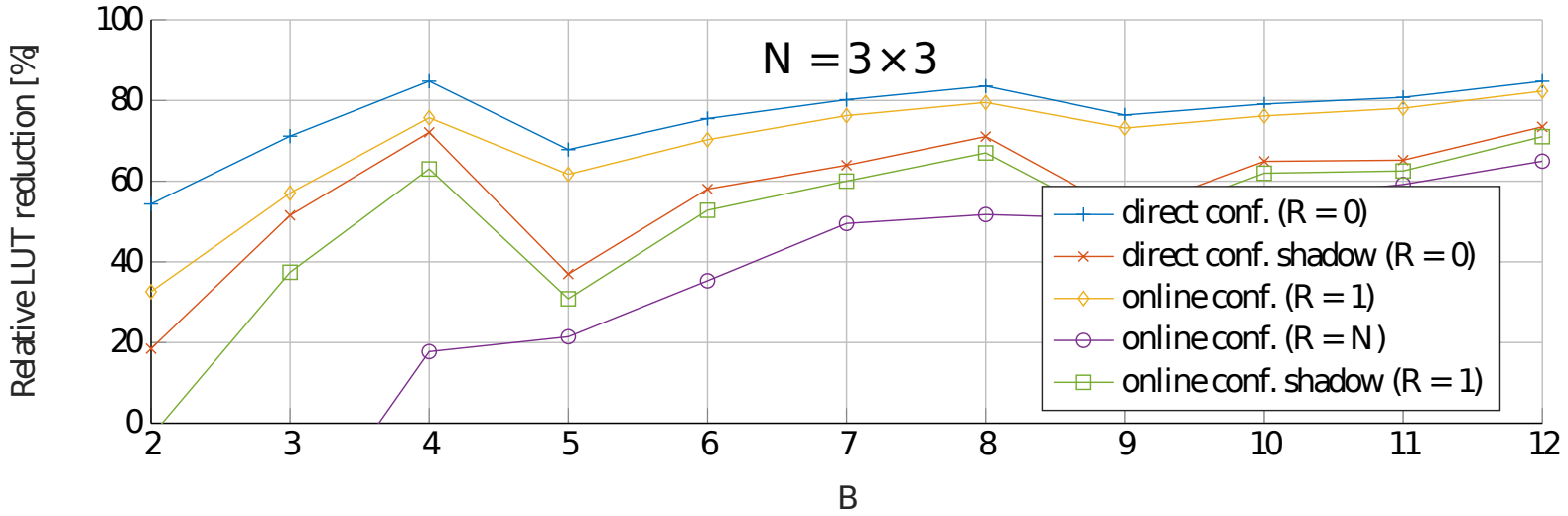
Synthesis results of conventional reference designs and several variants of the proposed design

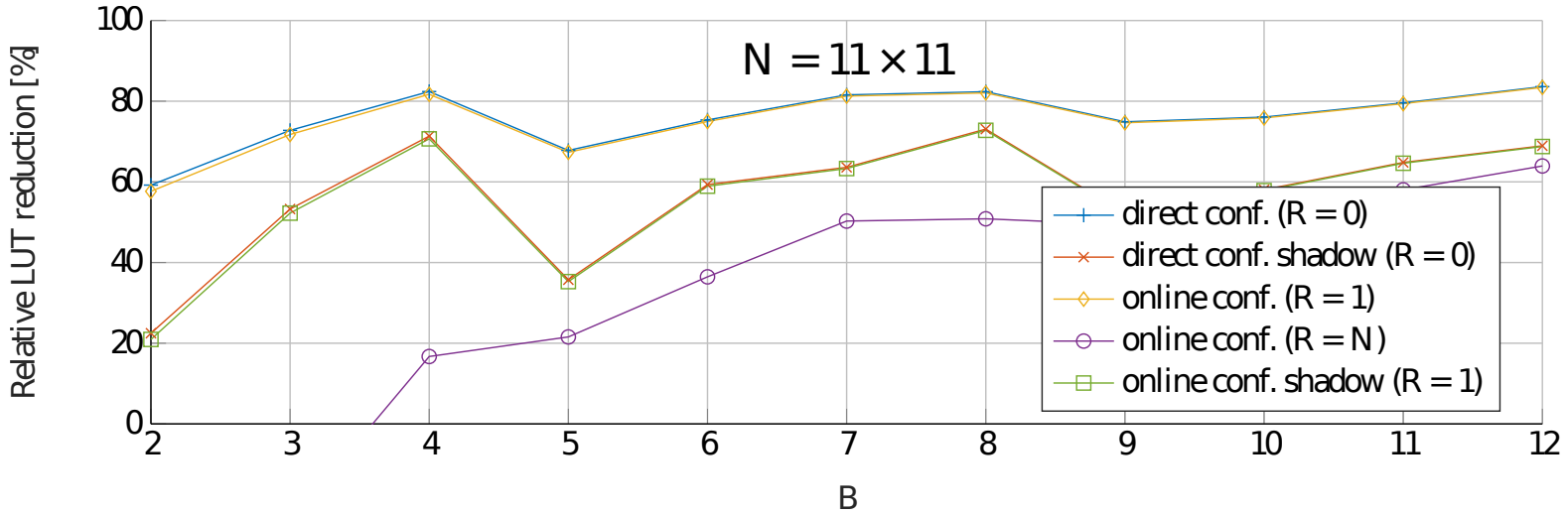
Kernel size (N)	Word size (B)	conventional SOP							proposed												
		DSP-based			logic only				direct conf.					online conf.							
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
3 × 3	2	0	108	487	92	751	18	7	42	517	75	467	576	62	517	182	517	95	467	18	4
3 × 3	3	0	213	363	163	480	27	7	47	485	79	474	576	70	485	214	485	102	474	27	4
3 × 3	4	0	334	337	276	661	36	7	42	410	77	423	576	67	410	227	410	102	423	36	4
3 × 3	5	5	344	312	457	432	45	7	147	311	288	281	2304	175	311	359	311	316	281	45	5
3 × 3	6	5	428	298	572	437	54	7	140	306	240	249	1728	170	306	370	306	270	249	54	5
3 × 3	7	9	81	379	838	383	63	7	166	319	302	257	2304	199	319	423	319	335	257	63	5
3 × 3	8	9	91	309	864	382	72	7	142	301	250	256	1728	177	301	417	301	285	256	72	5
3 × 3	9	9	99	429	1174	373	81	7	277	271	561	262	3456	315	271	579	271	599	262	81	6
3 × 3	10	9	110	355	1360	325	90	7	284	318	477	259	2880	324	318	604	318	517	259	90	6
3 × 3	11	9	122	424	1607	372	99	7	309	301	559	253	3456	352	301	656	301	602	253	99	6
3 × 3	12	9	131	434	1844	345	108	7	281	317	489	235	2880	326	317	646	317	534	235	108	6
11 × 11	2	0	1487	439	1299	629	242	10	530	312	1007	262	7744	550	312	2350	312	1027	262	242	7
11 × 11	3	0	2982	328	2247	421	363	10	612	324	1050	256	7744	635	324	2795	324	1073	256	363	7
11 × 11	4	0	4699	310	3692	576	484	10	650	300	1058	240	7744	675	300	3075	300	1083	240	484	7
11 × 11	5	61	5036	305	6030	419	605	10	1942	294	3875	243	30976	1970	294	4730	294	3903	243	605	10
11 × 11	6	61	6218	288	7790	408	726	10	1920	304	3169	254	23232	1950	304	4950	304	3199	254	726	10
11 × 11	7	121	1030	343	10845	376	847	10	1999	290	3947	262	30976	2032	290	5392	290	3980	262	847	10
11 × 11	8	121	1152	309	11536	356	968	10	2036	296	3105	117	23232	2071	296	5671	296	3140	117	968	10
11 × 11	9	121	1272	260	15670	354	1089	10	3935	270	7490	116	46464	3973	270	7933	270	7528	116	1089	11
11 × 11	10	121	1395	264	18109	345	1210	10	4335	261	7595	219	46464	4375	261	8575	261	7635	219	1210	11
11 × 11	11	121	1519	275	21352	342	1331	10	4358	271	7512	128	46464	4401	271	8961	271	7555	128	1331	11
11 × 11	12	121	1640	263	24627	327	1452	10	4036	253	7648	219	46464	4081	253	8881	253	7693	219	1452	11

Results: Convolution Core LUT comparison



Results: Convolution Core Reletive LUT Reduction





Conclusen

Result

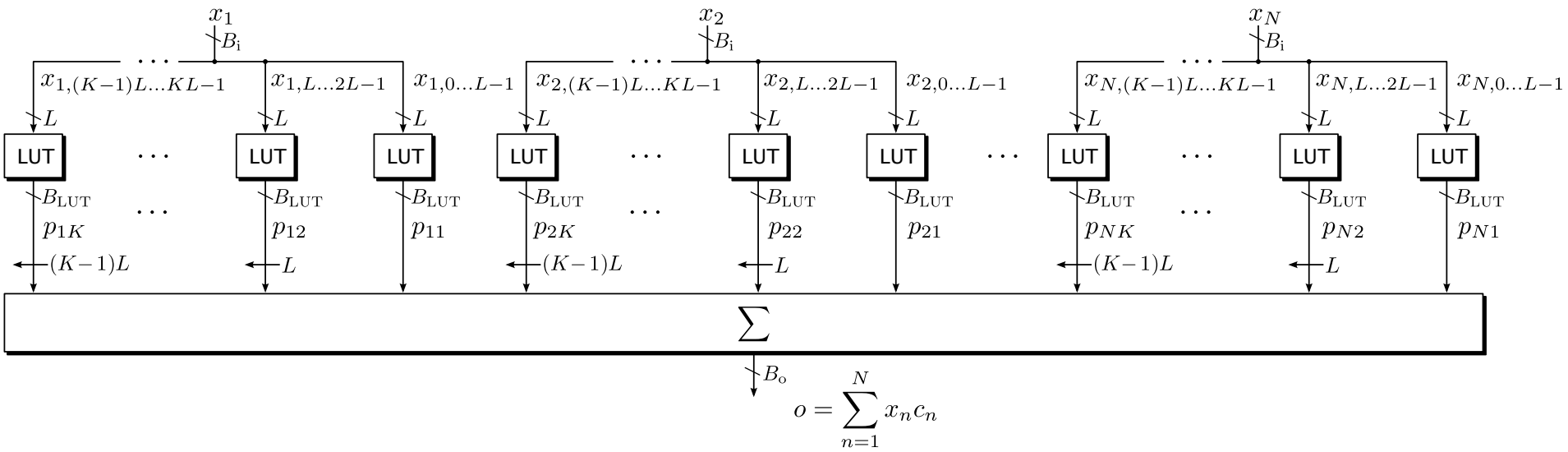
- Generator for plug in conv. kernals
- Open accacable in FloPoCo branch: uni_ks
- Adjustable in size and reconfiguration time
- Recource reduction up to 80%
- Recource reduction with shadow LUTs still 50%

Future Work

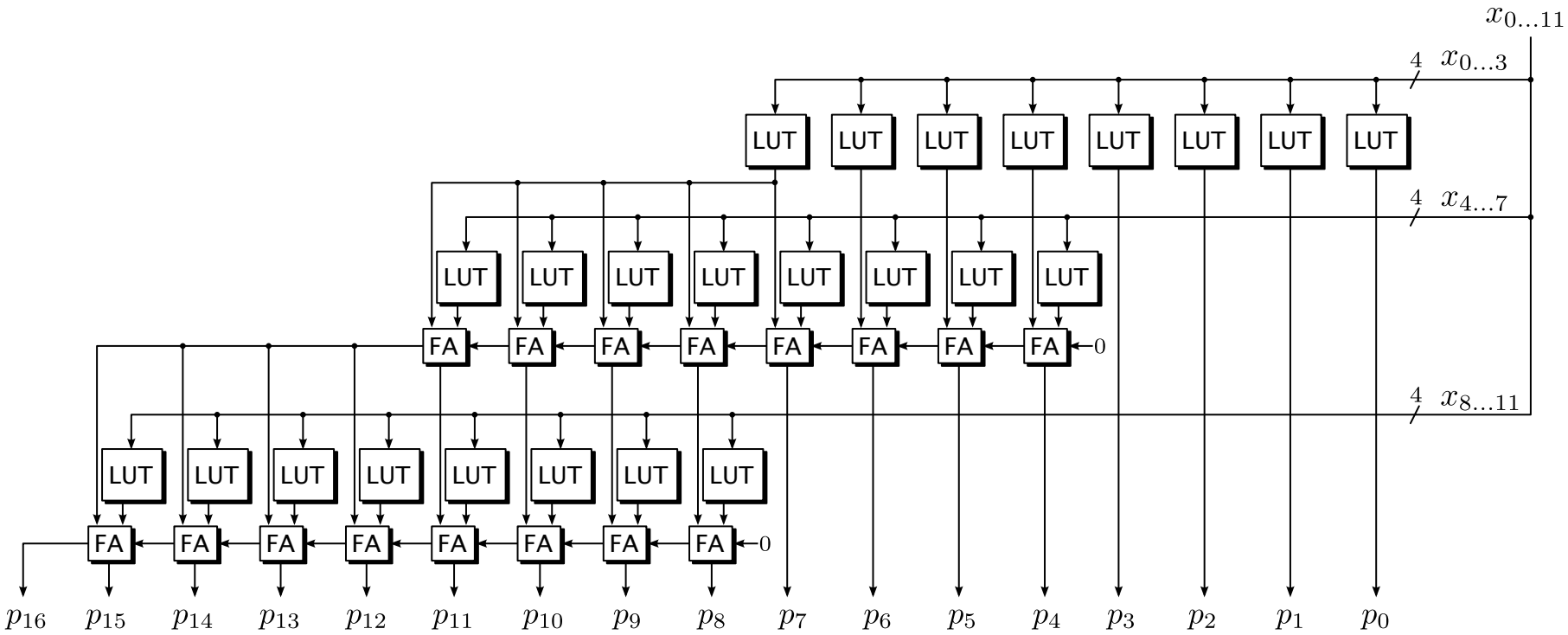
- Kernal test in working accelerators with complete network
- Complete network generator
- Heuristic to choose core parameters for a given Network

THANK YOU

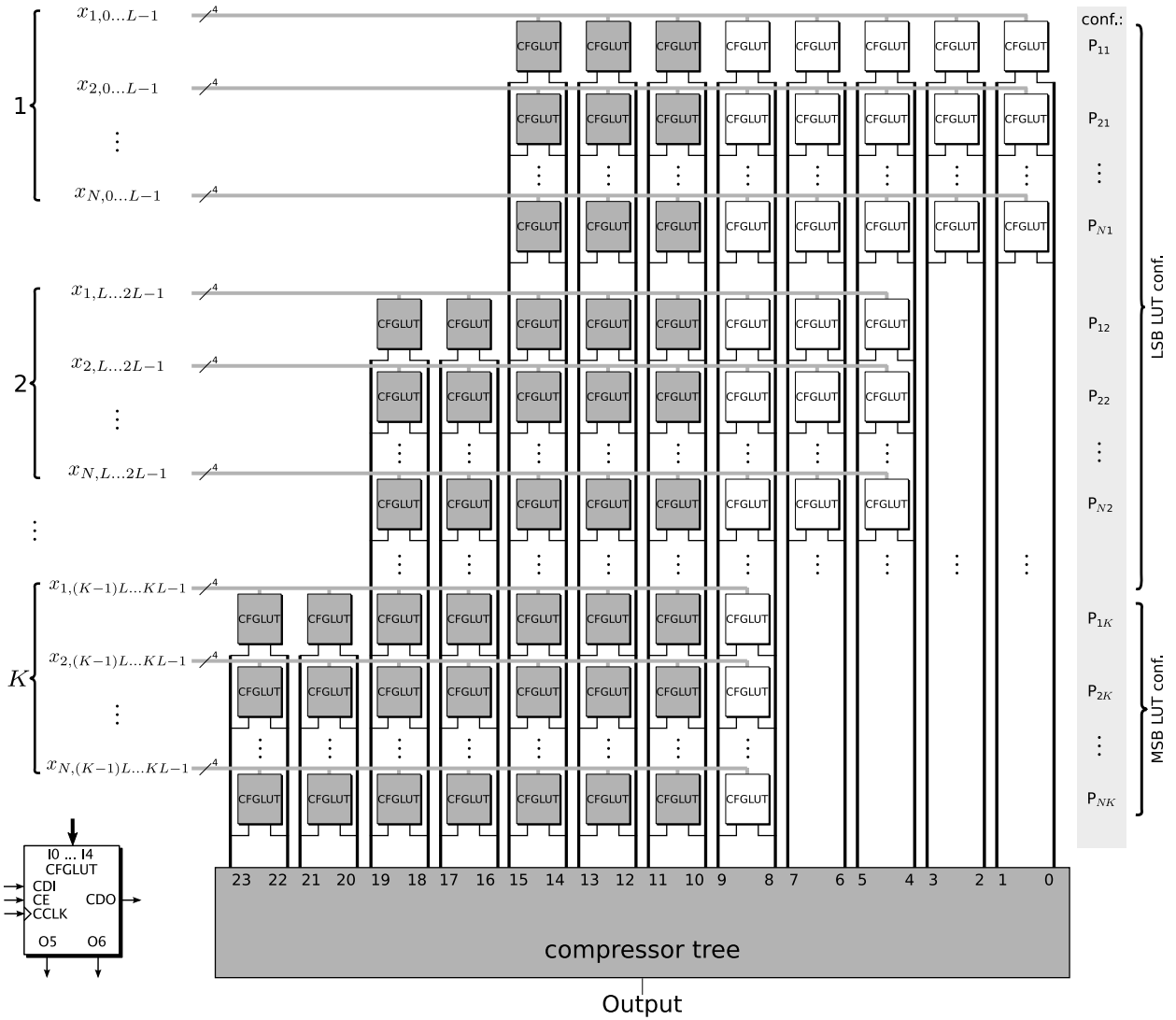
Generic LUT-based SOP



Generic LUT-based SOP



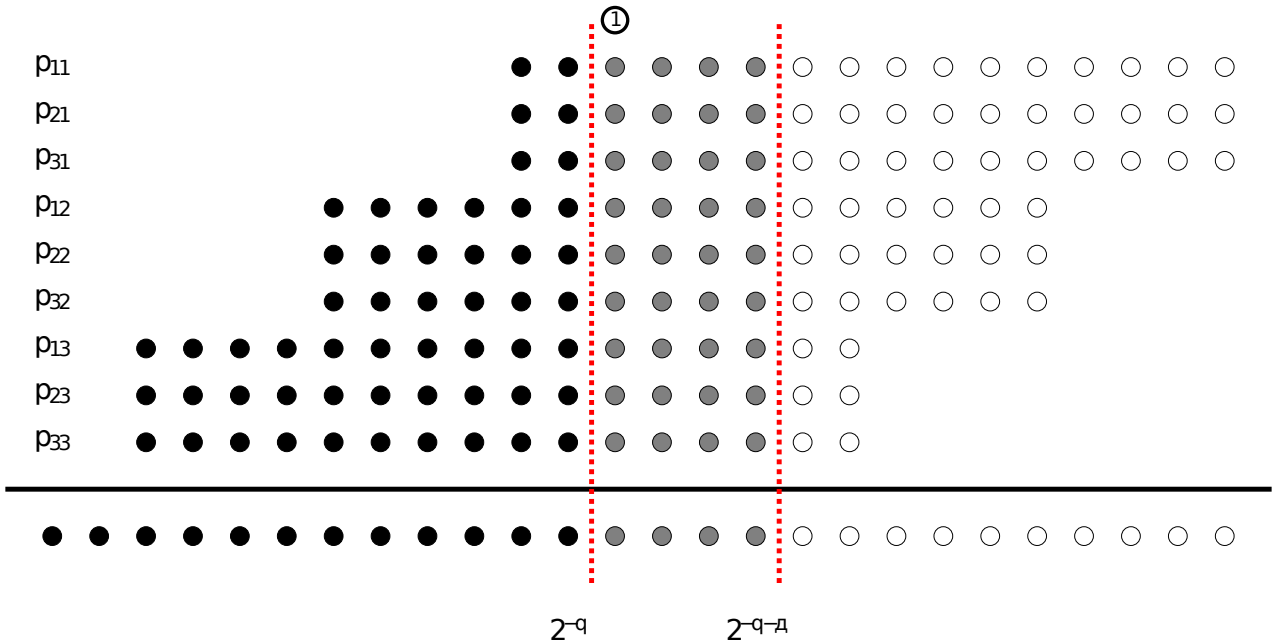
Structure for Convolutional Core



Faithful Rounding

Example

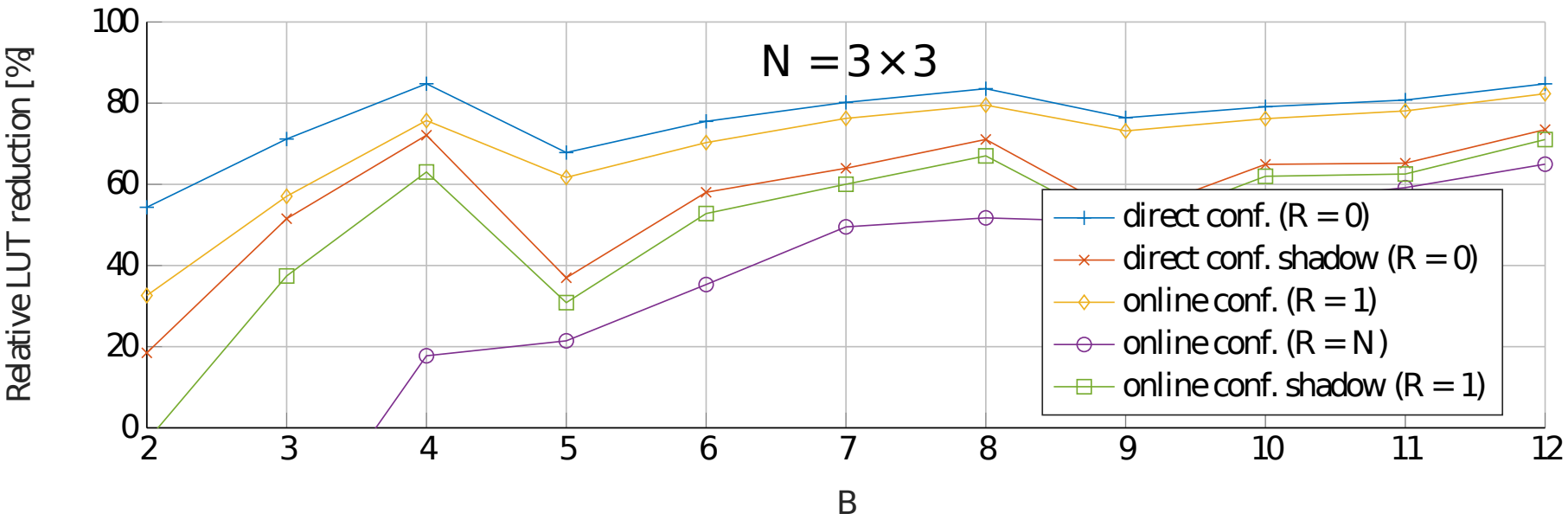
- SOP KCM of three 12 bit numbers
- 12 bit result
- 4 guard bits



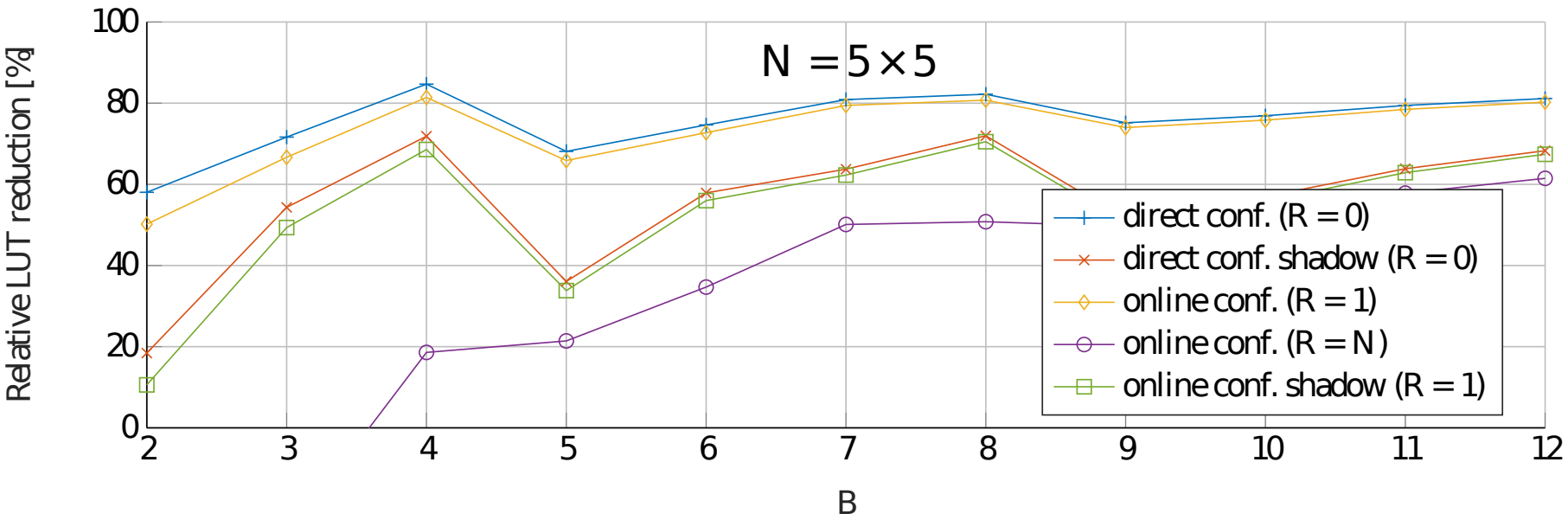
Configuration Time Overhead

Layer	N	Identical Ops (O)	R_{min}	Time Overhead Reconf. [%]		LUTs for $B = 8$ bit		
				$T_{rec} = 32$	$T_{rec} = 32N$	ref. (logic only)	online conf. shad. LUTs ($R = R_{min}$)	LUT red.
DarkNet19								
1	3×3	64,516	1	0.0	0.4	864	285	67.0%
2	3×3	16,384	1	0.2	1.8	864	285	67.0%
3,5	3×3	4,096	1	0.8	7.0	864	285	67.0%
4	1×1	4,096	1	0.8	0.8	77	65	15.6%
6	3×3	1,024	1	3.1	28.1	864	285	67.0%
7	1×1	1,024	1	3.1	3.1	77	65	15.6%
8	3×3	1,024	1	3.1	28.1	864	285	67.0%
9,11,13	3×3	256	2	12.5	112.5	864	315	63.5%
10,12	1×1	256	1	12.5	12.5	77	65	15.6%
14,16,18	3×3	64	5	50.0	450.0	864	405	53.1%
15,17,19	1×1	64	1	50.0	50.0	77	65	15.6%

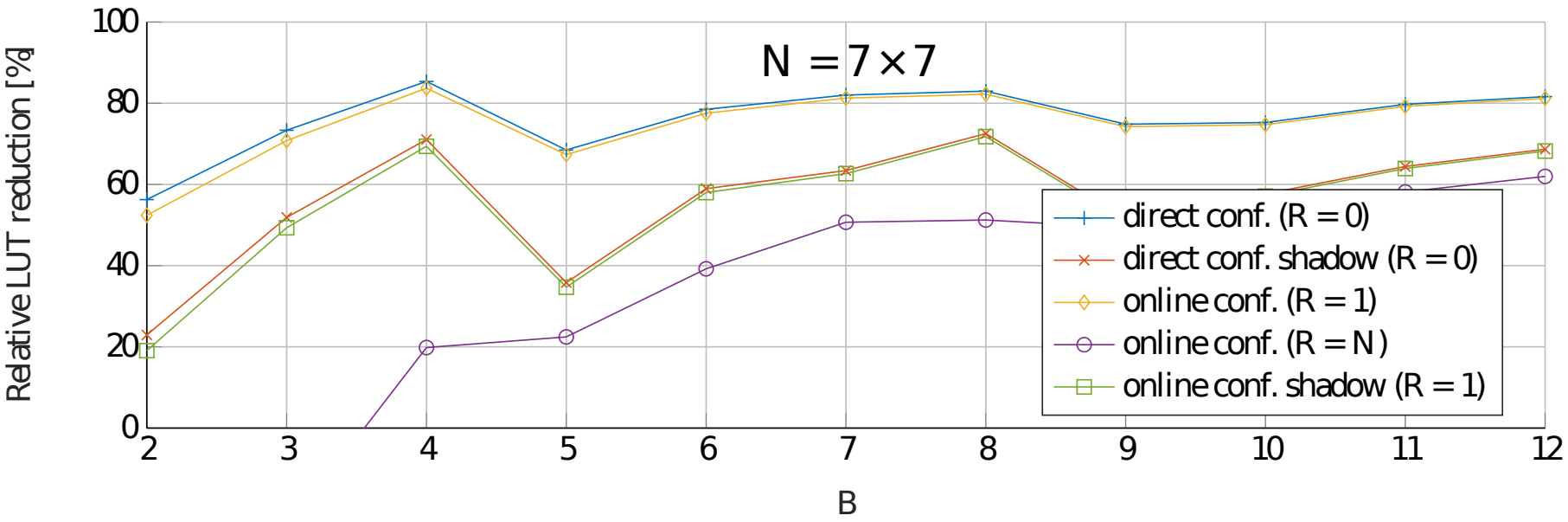
Results: Convolution Core Reletive LUT Reduction



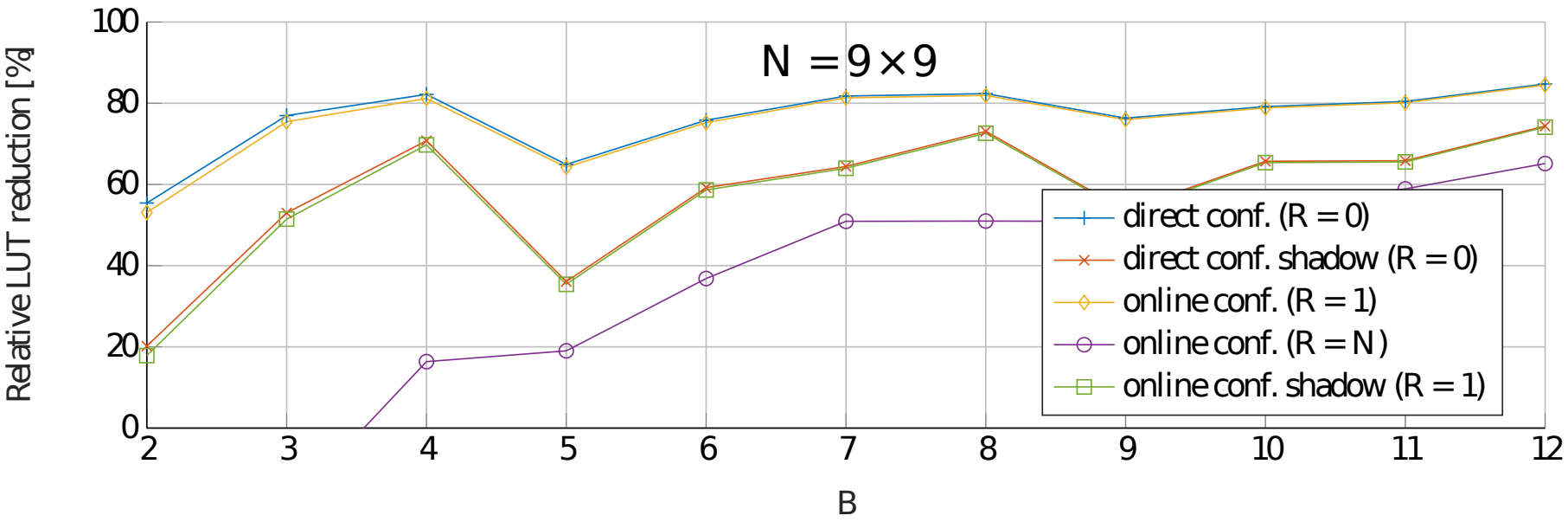
Results: Convolution Core Reletive LUT Reduction



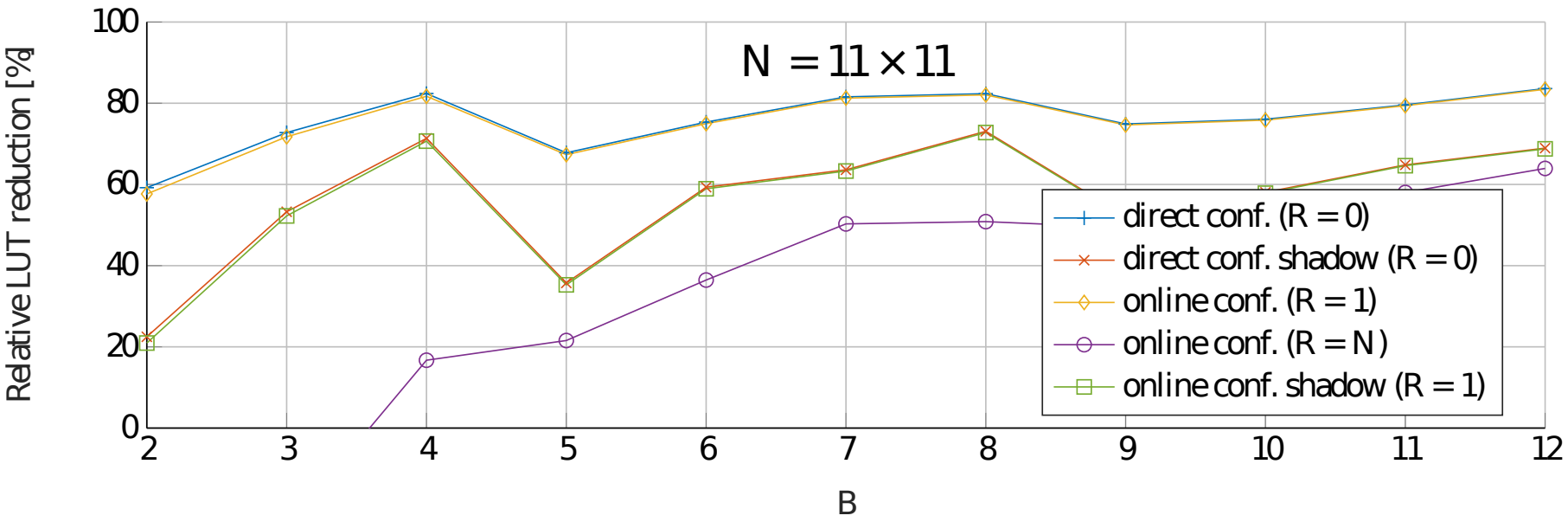
Results: Convolution Core Reletive LUT Reduction



Results: Convolution Core Reletive LUT Reduction



Results: Convolution Core Reletive LUT Reduction



Synthesis Results of Proposed Design 3x3

Synthesis results of conventional reference designs and several variants of the proposed design

Kernel size (N)	Word size (B)	conventional SOP							proposed											
		DSP-based			logic only				direct conf.					online conf.						
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
3 × 3	2	0	108	487	92	751	18	7	42	517	75	467	576	62	517	182	517	95	467	18	4
3 × 3	3	0	213	363	163	480	27	7	47	485	79	474	576	70	485	214	485	102	474	27	4
3 × 3	4	0	334	337	276	661	36	7	42	410	77	423	576	67	410	227	410	102	423	36	4
3 × 3	5	5	344	312	457	432	45	7	147	311	288	281	2304	175	311	359	311	316	281	45	5
3 × 3	6	5	428	298	572	437	54	7	140	306	240	249	1728	170	306	370	306	270	249	54	5
3 × 3	7	9	81	379	838	383	63	7	166	319	302	257	2304	199	319	423	319	335	257	63	5
3 × 3	8	9	91	309	864	382	72	7	142	301	250	256	1728	177	301	417	301	285	256	72	5
3 × 3	9	9	99	429	1174	373	81	7	277	271	561	262	3456	315	271	579	271	599	262	81	6
3 × 3	10	9	110	355	1360	325	90	7	284	318	477	259	2880	324	318	604	318	517	259	90	6
3 × 3	11	9	122	424	1607	372	99	7	309	301	559	253	3456	352	301	656	301	602	253	99	6
3 × 3	12	9	131	434	1844	345	108	7	281	317	489	235	2880	326	317	646	317	534	235	108	6

Synthesis Results of Proposed Design 5x5

Synthesis results of conventional reference designs and several variants of the proposed design

Kernel size (N)	Word size (B)	conventional SOP							proposed												
		DSP-based			logic only				direct conf.					online conf.							
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
5 × 5	2	0	346	469	255	701	50	8	107	323	208	267	1600	127	323	487	323	228	267	50	5
5 × 5	3	0	614	347	462	474	75	8	131	327	211	282	1600	154	327	586	327	234	282	75	5
5 × 5	4	0	963	308	764	652	100	8	117	302	215	281	1600	142	302	622	302	240	281	100	5
5 × 5	5	13	1015	311	1242	434	125	8	396	284	794	243	6400	424	284	976	284	822	243	125	7
5 × 5	6	13	1256	295	1577	402	150	8	400	306	664	236	4800	430	306	1030	306	694	236	150	7
5 × 5	7	25	216	497	2291	376	175	8	438	290	831	261	6400	471	290	1143	290	864	261	175	7
5 × 5	8	25	242	409	2404	357	200	8	428	311	674	239	4800	463	311	1183	311	709	239	200	7
5 × 5	9	25	266	242	3253	364	225	8	808	238	1577	101	9600	846	238	1638	238	1615	101	225	8
5 × 5	10	25	293	342	3739	340	250	8	864	302	1609	225	9600	904	302	1744	302	1649	225	250	8
5 × 5	11	25	321	365	4420	348	275	8	909	251	1598	120	9600	952	251	1864	251	1641	120	275	8
5 × 5	12	25	346	368	5108	334	300	8	963	269	1620	228	9600	1008	269	1968	269	1665	228	300	8

Synthesis Results of Proposed Design 7x7

Synthesis results of conventional reference designs and several variants of the proposed design

Kernel size (N)	Word size (B)	conventional SOP							proposed												
		DSP-based			logic only				direct conf.					online conf.							
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
7×7	2	0	665	441	519	671	98	9	227	323	400	267	3136	247	323	967	323	420	267	98	6
7×7	3	0	1205	311	904	438	147	9	241	325	435	282	3136	264	325	1128	325	458	282	147	6
7×7	4	0	1924	321	1503	602	196	9	220	305	435	252	3136	245	305	1205	305	460	252	196	6
7×7	5	25	2022	304	2459	420	245	9	775	271	1578	240	12544	803	271	1907	271	1606	240	245	7
7×7	6	25	2466	284	3131	419	294	9	673	279	1285	249	9408	703	279	1903	279	1315	249	294	7
7×7	7	49	420	435	4397	370	343	9	791	244	1609	256	12544	824	244	2168	244	1642	256	343	7
7×7	8	49	470	384	4647	366	392	9	791	290	1277	114	9408	826	290	2266	290	1312	114	392	7
7×7	9	49	518	309	6369	357	441	9	1603	294	3086	249	18816	1641	294	3225	294	3124	249	441	9
7×7	10	49	569	409	7365	354	490	9	1823	306	3118	229	18816	1863	306	3543	306	3158	229	490	9
7×7	11	49	621	308	8651	343	539	9	1754	303	3078	116	18816	1797	303	3621	303	3121	116	539	9
7×7	12	49	670	324	10001	334	588	9	1838	304	3133	241	18816	1883	304	3803	304	3178	241	588	9

Synthesis Results of Proposed Design 9x9

Synthesis results of conventional reference designs and several variants of the proposed design

Kernel size (N)	Word size (B)	conventional SOP							proposed												
		DSP-based			logic only				direct conf.					online conf.							
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
9 × 9	2	0	1100	436	857	609	162	10	382	315	684	268	5184	402	315	1602	315	704	268	162	6
9 × 9	3	0	2025	333	1506	444	243	10	347	267	708	263	5184	370	267	1810	267	731	263	243	6
9 × 9	4	0	3198	318	2469	598	324	10	440	292	722	266	5184	465	292	2065	292	747	266	324	6
9 × 9	5	41	3363	306	4070	424	405	10	1428	290	2601	235	20736	1456	290	3296	290	2629	235	405	9
9 × 9	6	41	4157	285	5208	418	486	10	1260	295	2124	263	15552	1290	295	3290	295	2154	263	486	9
9 × 9	7	81	698	262	7367	375	567	10	1343	279	2621	262	20736	1376	279	3616	279	2654	262	567	9
9 × 9	8	81	780	308	7760	362	648	10	1369	272	2090	131	15552	1404	272	3804	272	2125	131	648	9
9 × 9	9	81	860	245	10528	307	729	10	2492	249	4879	115	31104	2530	249	5170	249	4917	115	729	10
9 × 9	10	81	943	330	12095	346	810	10	2520	287	4148	130	25920	2560	287	5360	287	4188	130	810	10
9 × 9	11	81	1031	369	14321	340	891	10	2802	303	4892	117	31104	2845	303	5885	303	4935	117	891	10
9 × 9	12	81	1112	345	16553	331	972	10	2524	273	4241	226	25920	2569	273	5769	273	4286	226	972	10

Synthesis Results of Proposed Design 11x11

Synthesis results of conventional reference designs and several variants of the proposed design

Kernel size (N)	Word size (B)	conventional SOP							proposed												
		DSP-based			logic only				direct conf.					online conf.							
									R=0		shadow R=0			R=1		R=N		shadow R=1			
		DSPs	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	LUTs	f_{max}	LUTs	f_{max}	LUTs	f_{max}	M_{kern}	Latency
11 × 11	2	0	1487	439	1299	629	242	10	530	312	1007	262	7744	550	312	2350	312	1027	262	242	7
11 × 11	3	0	2982	328	2247	421	363	10	612	324	1050	256	7744	635	324	2795	324	1073	256	363	7
11 × 11	4	0	4699	310	3692	576	484	10	650	300	1058	240	7744	675	300	3075	300	1083	240	484	7
11 × 11	5	61	5036	305	6030	419	605	10	1942	294	3875	243	30976	1970	294	4730	294	3903	243	605	10
11 × 11	6	61	6218	288	7790	408	726	10	1920	304	3169	254	23232	1950	304	4950	304	3199	254	726	10
11 × 11	7	121	1030	343	10845	376	847	10	1999	290	3947	262	30976	2032	290	5392	290	3980	262	847	10
11 × 11	8	121	1152	309	11536	356	968	10	2036	296	3105	117	23232	2071	296	5671	296	3140	117	968	10
11 × 11	9	121	1272	260	15670	354	1089	10	3935	270	7490	116	46464	3973	270	7933	270	7528	116	1089	11
11 × 11	10	121	1395	264	18109	345	1210	10	4335	261	7595	219	46464	4375	261	8575	261	7635	219	1210	11
11 × 11	11	121	1519	275	21352	342	1331	10	4358	271	7512	128	46464	4401	271	8961	271	7555	128	1331	11
11 × 11	12	121	1640	263	24627	327	1452	10	4036	253	7648	219	46464	4081	253	8881	253	7693	219	1452	11