Physical Synthesis for Advanced Neural Network Processors

Zhuolun He[1], Peiyu Liao[1], Siting Liu[1], Yuzhe Ma[1], Yibo Lin[2], Bei Yu[1],

[1]The Chinese University of Hong Kong
[2]Peking University

Outline

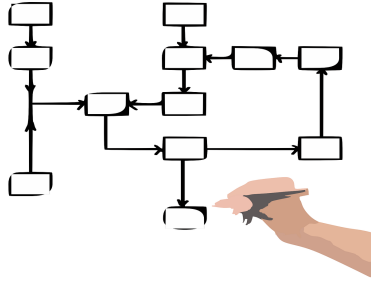
Outline

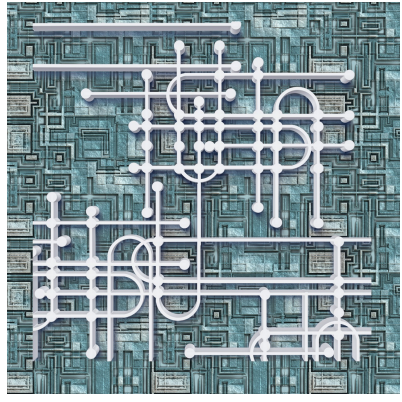
Deep Learning: an Important Workload

- Deep learning has emerged as an important workload
- Yet, how to efficiently execute them?

Dataflow Optimization and Physical Synthesis

- Dataflow optimization becomes essential
  - Schedules operation by data availability
  - Exposes opportunity for parallelism and data reuse
- Dataflow regularity gives rise to new physical synthesis approaches
  - Directly determines system performance!





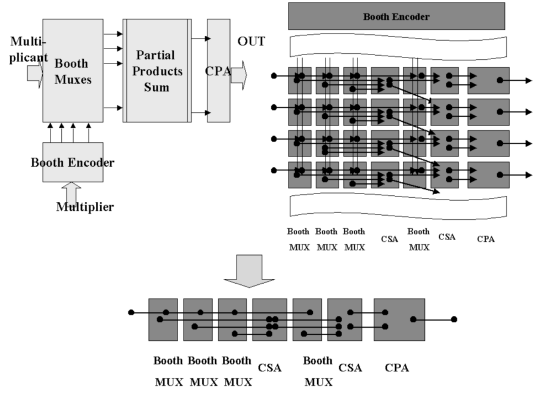
Outline

Datapath Driven Standard Cell Placement

- Classical idea: bit-sliced DSP datapaths [Cai, DAC'90]
  - Decide ordering of linearly placed blocks
  - Solved by A$*$ in the search space
- Standard cell placement [Tsay, TCAD'95]
  - Strongly connected subcircuits (cones) are extracted
  - BFS + heuristics
  - Placed as macro cells

Datapath Driven Placement: Abstract Physical Model

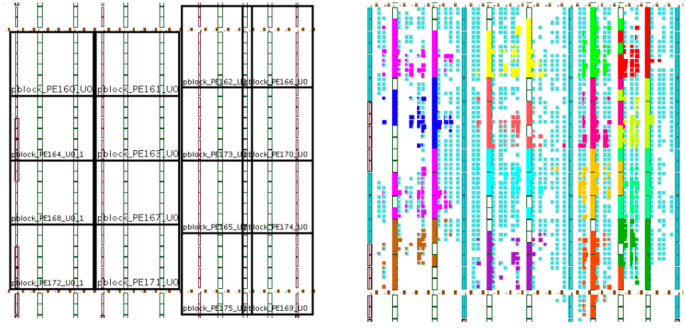
- Placer has little control of exact locations if datapath is generated separately
- Abstract physical model [Ye, ICCAD'00]
  - Bit-sliced abstraction
  - Compiled from HDL
  - Blocks are placed abutted
- Two-step heuristic for linear placement
  - quadratic assignment
  - sliding window optimization



APM of a booth multiplier [Ye, ICCAD'00]

Datapath Driven Systolic Array Placement

- Systolic arrays are a popular choice to support neural network computations
- Current FPGA CAD tools cannot synthesize them in high quality
- One solution: restrict fixed locations for PEs [Zhang, ISCAS'19]
  - Sufficient DSPs, close to used I/O banks



PE placement with floorplan constraints [Zhang, ISCAS'19]

Datapath Driven Placement: Many More

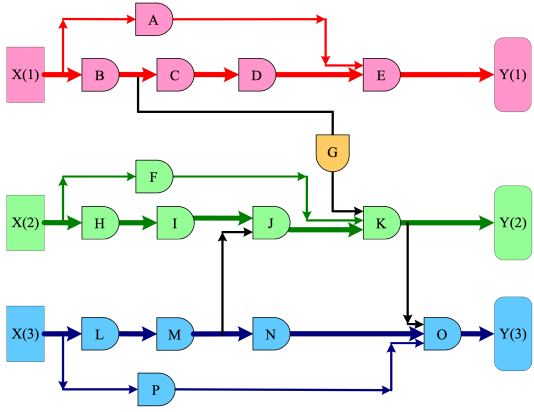
- Detailed placement [Serdar, DATE'01]
- SOC placement [Tong, JOS'02] [Jing, ICCCAS'02]
- Parallel multiplier design [Bae, ISPD'15]
- Genaral ASIC design [Ye, ISCAS'02] [Chou, DAC'12] [Wang, IETCDS'17]
- . . .

Regularity Extraction

- Consider cells with the same bit-slice are lined up horizontally [Nijssen IWLS'96]
  - geometric regularity: circuit is fitted onto a matrix of rectangular buckets
  - interconnect regularity: most nets are within one slice/one stage
- Typical methods for regularity extraction
  - Search-wave expansion [Nijssen IWLS'96]
  - Signature-based [Arikati, ICCAD'97]
  - Template covering [Chowdhary, TCAD'99]
  - Network flow [Xiang, ISPD'13]
  - Bipartite graph vertex covering [Huang, DAC'17]

Regularity Extraction: Network Flow Approach

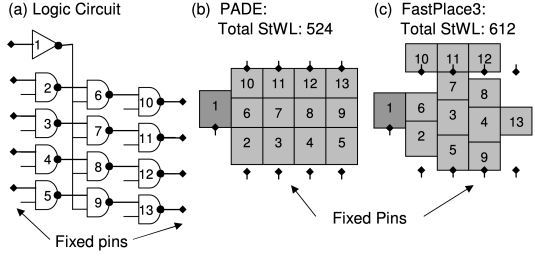
- *Datapath main frame* (DMF) [Xiang, ISPD'13]
  - a set of n disjoint paths from input to output
  - maximize the number of datapath gates on these paths
- Can be optimally identified by the min-cost max-flow algorithm



DMF identification can be transformed to a network flow problem [Xiang, ISPD'13]

Machine Learning Techniques for Datapath Driven Placeme

- Machine learning techniques are involved [Ward, DAC'12]
  - Graph-based (e.g., automorphism) and physical features (e.g., cell area) are analyzed and extracted from the netlist
  - Features are fed to SVMs and NNs to classify and evaluate datapath patterns
  - Maximize the evaluation accuracies of datapath and non-datapath pattern
  - Proposed new placer: PADE

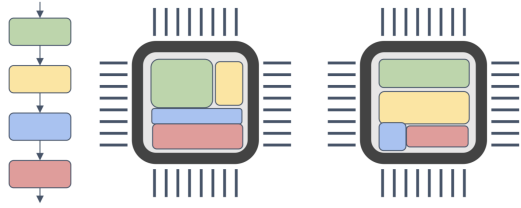


PADE effectively handles datapath in placement. [Ward, DAC'12]

Outline

Problem Formulation

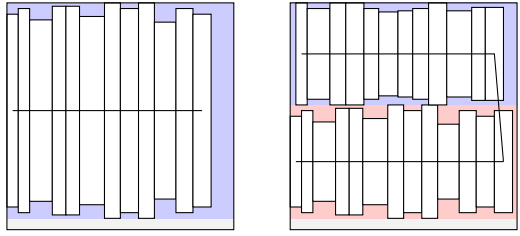
- Floorplan computation kernels of a neural network [Michael, ISPD'20]
  - Blocks are soft: kernel sizing
  - Floorplanning
  - Optimize performance, wirelength, etc.



Kernel floorplanning. Figure adopted from ISPD Contest.

Datapath Driven Floorplan

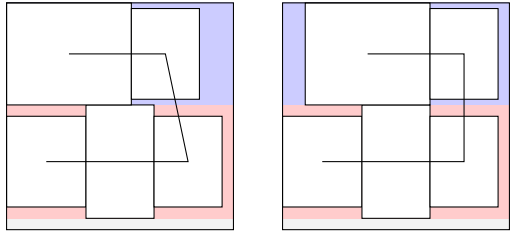
- Our strategy: datapath-driven floorplan
  - Neural network is usually hierarchical: a stack of layers
  - Arrange the floorplan according to the datapath



Datapath driven floorplans. (a) Single-row; (b) Multi-row.

Floorplan Compacting

- The floorplan can be further compacted
  - Vertically: push to center
  - Horizontally: a linear programming problem



(a) Mamba floorplan; (b) Horizontally compacted Mamba floorplan.

Result: Datapath Driven Floorplan Outperforms

Benchmark Statistics				TBS SA Algorithm [Jiang, ICCAD'20]				Our Strategy			
Case	#Kernels	λ_1	λ_2	Max Time	WL	Protocol	Cost	Max Time	WL	Protocol	Cost
A	17	1	0	37044	3611.5	11	40655.5	35280	2047	13	**37327**
B	34	1	0	70560	6657	20	77217	65856	4905	17	**70761**
C	102	1	0	76608	15696	69	92034	65772	4278	281	**70050**
D	54	1	0	38304	9327.5	44	47631.5	34944	3071.5	89	**38015.5**
E	17	10	100	36288	2080.5	7	57793	39690	590	16	**47190**
F	34	10	100	76608	3237	15	110478	70560	1475	14	**86710**
G	102	10	100	91728	7784	29	172468	69888	2508	141	**109068**
H	54	10	100	47040	4450	21	93640	43008	893	115	**63438**
I	27	4	0	56448	3790	16	71608	52920	612	13	**55368**
J	81	4	0	63504	8009.5	52	95542	57792	1117.5	286	**62262**
K	18	4	0	576	236	3	**1520**	504	400	14	2104
L	54	4	0	1280	910.5	60	4922	504	774	114	**3600**
M	25	4	0	2359296	9359	24	2396732	2336256	5100	67	**2356656**
N	28	4	0	2268	707.5	0	5098	1599	448.5	9	**3393**
O	27	40	400	63504	1202	6	113984	52920	612	13	**82600**
P	81	40	400	115101	4015	24	285301	66528	2273	102	**198248**
Q	18	40	400	1152	178	1	**8672**	504	400	14	22104
R	54	40	400	1372	1443	30	**71092**	504	774	114	77064
S	25	40	400	2495376	3551	25	2647416	2396160	1899.5	65	**2498140**
T	28	40	400	5720	555.5	0	27940	2015	367.5	9	**20315**
Avg							1.17×				**1.00**×

Outline

Processing-in-Memory Technologies

- PIM provides massive parallelism with high energy efficiency [Wang, ASAP'19]
- However, lots of systems rely on external control
- Neural networks can be implemented with NVMs
  - RRAM [Chi, SIGARCH'16] [Sun, DATE'18]
  - STT-MRAM [Yan, SC'18] [Pan, TMAGN'18]
  - PCM [Kim, DATE'20]
  - memristor [Yao, Nature'20]
- In-memory analog simulation
  - memristor crossbar [Li, Nat. Commun'18]
  - FTJ [Li, Adv.Mater'20]

Nanophotonic and Optical Neural Networks

- The nanophotonic circuit is an alternative neuromorphic computing system
  - ultra-high bandwidth, speed and ultra-low energy consumption
- Signals are encoded in the amplitude of optical pulses
- Implements a multi-layer optical neural network (ONNs)
- Recent advances
  - Exploring nonlinear functions with ONNs [Zuo, OPTICA'19]
  - Reducing area overhead of ONNs [Gu, ASPDAC'20]

3D Technologies and Beyond-CMOS devices

- 3D technologies offer a wide spectrum of integration schemes
  - TSV-based Tetris [Gao, ASPLOS'17]
  - TCI-based QUEST [Ueyoshi, ISSCC'18]
  - M3D-based accelerators [Chang, ISLPED'17] [Chang, JETC'18]
- Beyond-CMOS devices give rise to new solutions for low-power design
  - HyperFET for spiking neural networks [Tsai, TMSCS'16]
  - TFET for cellular neural networks [Palit, ISLPED'13]

Outline

Conclusion

- Review on datapath driven placement
  - Placement with datapath constraints
  - Regularity extraction methods
  - Machine learning techniques
- Case study: floorplan for advanced neural network processors
  - Datapath driven floorplan greatly outperforms standard methods
- Advanced technologies for neural network processor design
  - Processing-in-Memory
  - Nanophotonic
  - 3D technologies
  - Beyond-CMOS devies

Thank You