

A 3DIC system to aid in the acceleration of systems that employ multiple instances of artificial neural networks

Final exam : Lee B. Baker

Advisor: Paul Franzon

Outline

1. What are artificial Neural networks
2. Target application and ANN type's
3. State-of-the-art
4. System Architecture
 - Problem
 - Solution
5. System details
 - 3DIC DRAM and data structures
 - 3D Bus
 - Manager
 - Processing engine
6. Results
7. Summary

Introduction

- Recently Artificial Neural networks (NN) have demonstrated superior performance in classification and function approximation

Deep Neural Networks (DNN) have been very successful in image processing applications [Kri12]

- it is anticipated that DNNs will be employed more and more in self-driving cars



Reinforcement learning is gaining popularity

- alphaGo employed reinforcement learning with deep neural networks [Mad14]

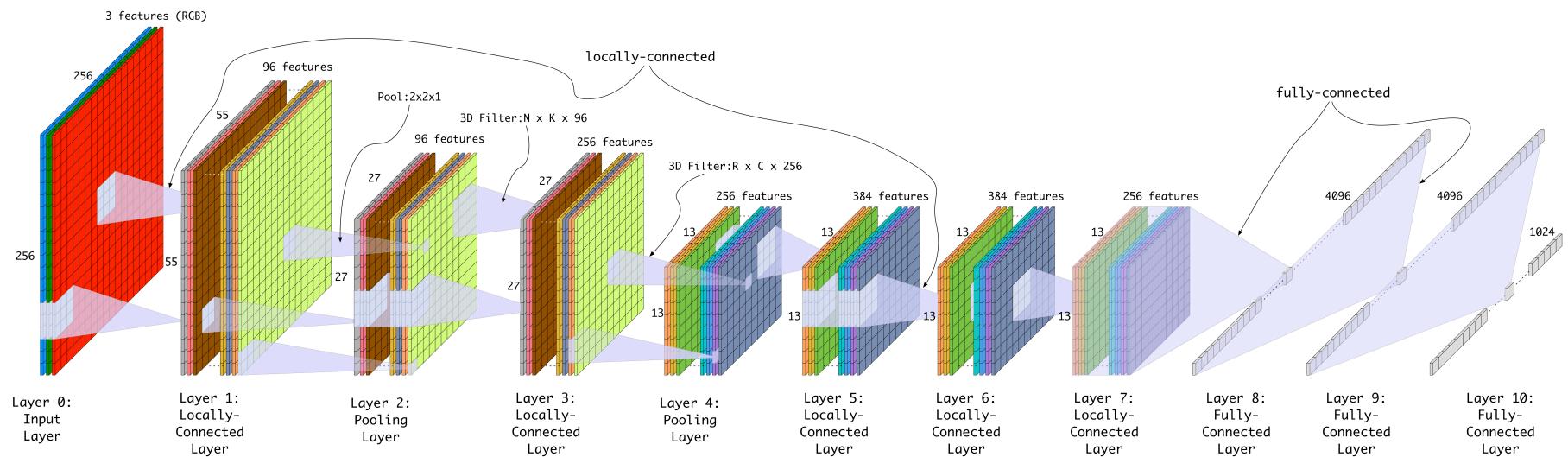


Introduction

- What about applications that will and do require multiple NN processors running at or near real-time?
 - drone scanning in many directions during navigation
 - looking down different than looking forward??
 - aircraft performing system diagnostics during flight
 - observing thermal profiles
 - observing vibrations
 - airport security
 - face recognition
 - body thermal profiles
 - motion

Introduction

- Most useful ANNs are big
 - 100's of thousands of artificial neurons
 - real-time processing requires large amounts of memory bandwidth and storage
 - DRAM is required
 - Most implementations employ SRAM for local memory



Research

- Most research focused on CNNs
 - can take advantage of SRAM using parameter reuse
- A lot of focus on server applications
 - perhaps lots of research \$\$ available??
 - server applications can take advantage of SRAM for batch processing
- But many DNNs are fully-connected [XX]
 - LSTM and MLPs
 - classifier stage of all ANNs including CNN

Mission Statement

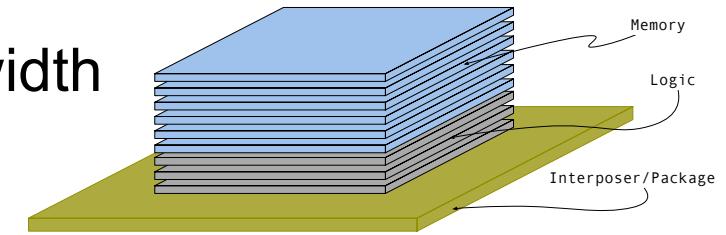
- Real world embedded applications will require multiple disparate NNs to be solved simultaneously
 - If ANNs fulfill their potential, they will be used for various system functions
- Need the capacity provided by DRAM
 - Avoid dependency on SRAM
 - SRAM is easy to use, but capacity limitations impose ANN size restrictions.
- Need to consider the system impact
 - Research often focuses on point problems. Need to consider interaction between blocks.

Problem

- Many embedded applications will/do have power, space and weight limitations
 - to achieve near or at real-time processing, current solutions require high power and high real-estate
 - GPU solutions large and high power
 - ASIC's better than GPUs
 - Both have memory bandwidth limitations when reuse opportunities do not exist
 - multiple disparate ANNs means limited reuse opportunities and thus SRAM is of limited use
 - no or limited batch processing

Solution

- 3DIC Architecture
 - reduces energy and area
 - increase connectivity and bandwidth
- 3D-DRAM
 - provide high bandwidth and large storage
 - operate directly out of DRAM
- Data Structures
 - data structures to ensure neural network data can be accessed efficiently
- Specialized processing layers
 - provide special functions to aid in acceleration of target neural networks



Feasibility

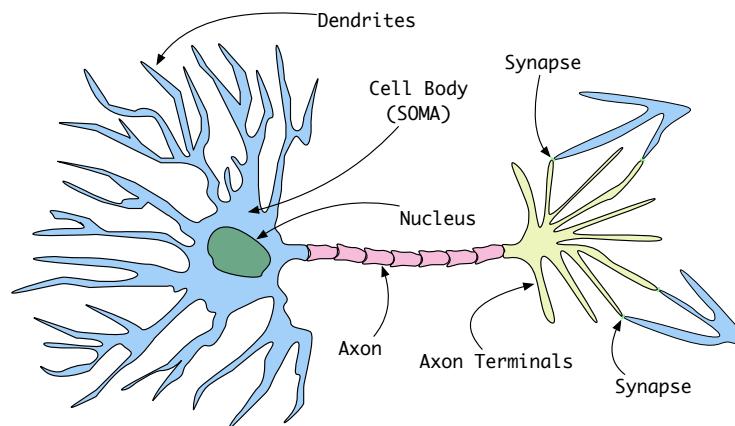
- Can a 3D-DRAM be used effectively?
- Can a useful system fit within the 3D stack footprint?
- How can we control such a system?

Contribution

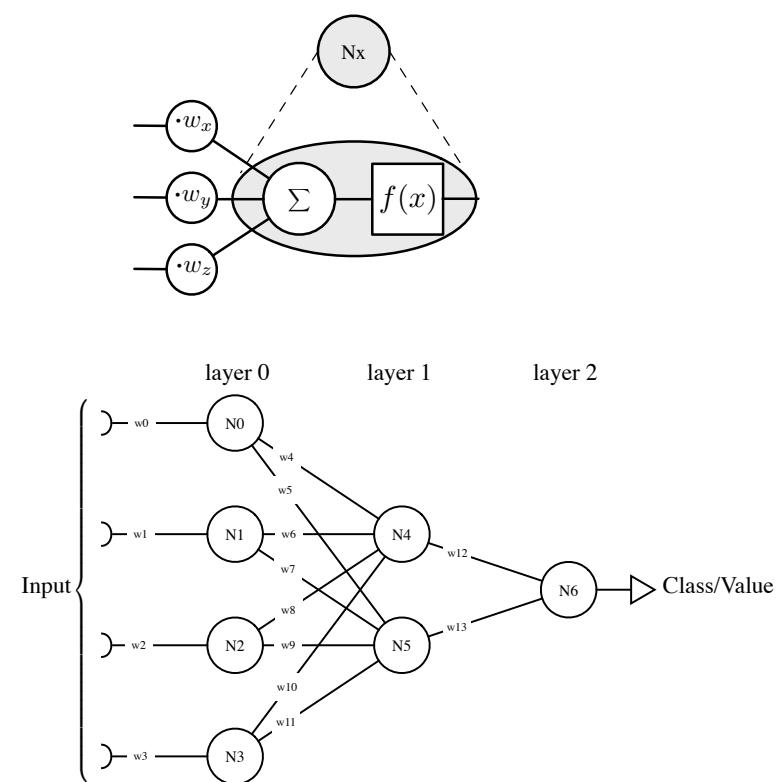
- An extensible architecture that can simultaneously process multiple disparate real-time DNNs
- Proposed a custom 3D-DRAM providing a ~32X bandwidth benefit compared to standard 3D-DRAM
- A DNN system solution that employs pure 3DIC technology
- Custom instructions and data structures that facilitate operating directly out of 3D-DRAM

Artificial Neural Networks

- a network of processing elements inspired by the connectivity and processing observed in the brain



- the processing elements or neurons are connected together to form a network



Artificial Neural Networks

- the processing elements fall into two categories
 - rate based neurons which try to capture the neuron behavior in the form of a number
 - relatively simple to model
 - easier to train and have shown high levels of efficacy in many applications
 - spiking neurons which more closely emulate actual brain behavior
 - model neural activity in the form of differential equations
 - require numerical methods to solve although there have been attempts to employ analog circuits
- In this work we are focusing on rate based ANNs

Summary : Target application and ANN type's

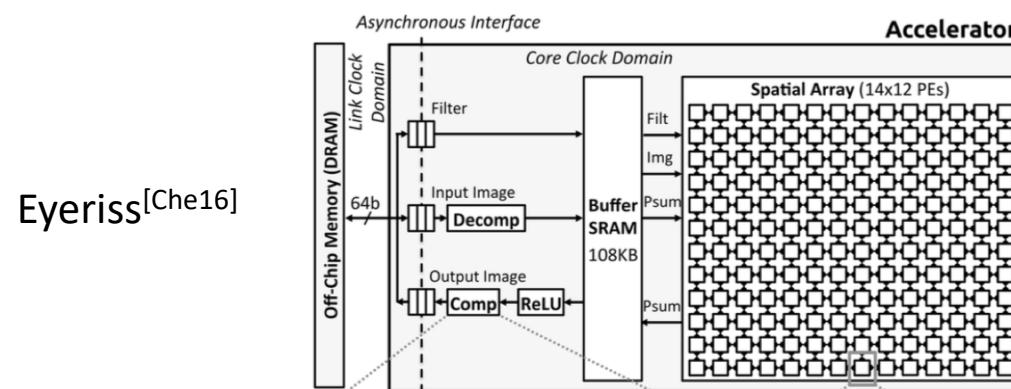
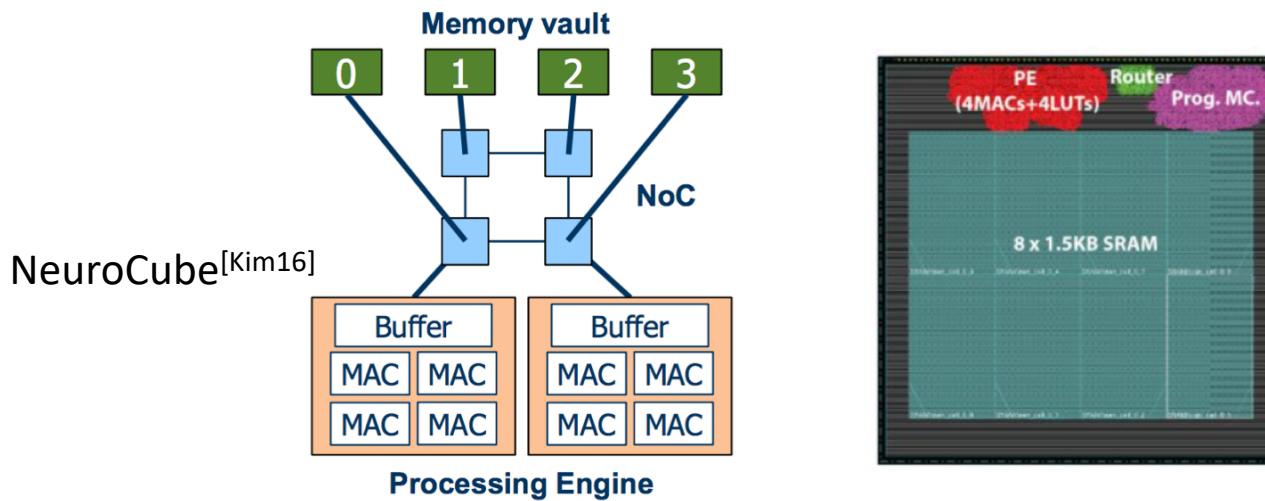
- Embedded systems
 - not cloud based
 - assumes multiple tasks are being performed
 - requires multiple disparate ANNs
 - inference only
- This work has focused on support for:
 - fully- and locally-connected DNNs
 - can be extended to support LSTMs

State-of-the-art

- Current implementations either use GPU's or low capacity ASIC implementations
- GPU's are typically general purpose devices and consume high power
 - GPU's do lend themselves to convolution when weights are shared
 - ASICs almost always provide performance improvements
- ASIC implementations target larger networks by using multiple devices
 - usually assume local SRAM which limits network size
 - some/most implementations target specific NN's, usually CNNs

State of the Art ASIC's

- ASIC's



State of the Art ASIC's

- NeuroStream
 - acknowledges DRAM is required for useful DNNs
 - uses HBM 3D-DRAM along with local SRAM
 - depends on SRAM for performance
- Google TPU
 - assumes large levels of batch processing
 - 8-bit processing
 - fully-connected NNs represent very high percentage
 - acknowledges performance degradation using fully-connected
- DaDiannao
 - uses embedded DRAM but still requires up to 64 for useful DNNs
 - high internal bandwidth but dissipates high power when scaling to useful-sized ANNs

¹ [Kim16], ² [Che16], ³ [Esm05], ⁴ [Far11]

State of the Art

- Current state-of-the art ASIC's and GPU's do not provide the capacity and bandwidth to provide a solution requiring multiple useful disparate ANN's to be computed in real-time
- Goldilocks principle
 - need balance of bandwidth and storage
 - need to read all parameters in sample time for real-time processing
 - Currently SOA bandwidth too low or too high

System Requirements

- System solution requiring multiple NN's
 - ~8GB of memory for baseline ANN
 - real-time processing : ~16mS (60 frames/sec)
 - ~26Tbps memory bandwidth for baseline ANN
- Current 3D-DRAM memory technology bandwidth
 - HMC – SERDES~2Tbps – too low
 - HBM – wide DDR ~2Tbps – too low
 - Tezzaron DiRAM4 – wide DDR – 4Tbps – too low
- Propose Customized DiRAM4
 - expose more of the page
 - take advantage of high density TSVs

Architecture Blocks

- Proposed Customizations to standard DiRAM4 3D-DRAM
- Data Structures and Instructions to support efficient use of DRAM
- Management Layer for configuration and control
 - DRAM controller
 - Instruction decoder
- Processing layer targeted toward DNNs
 - streaming functions operate directly on data
 - multiply/accumulates ANe activation
 - multiply for softmax
 - additional special functions in SIMD used in the neuron activation process
 - ReLu for ANe activation
 - divide and exponent for softmax function
 - compare for pooling

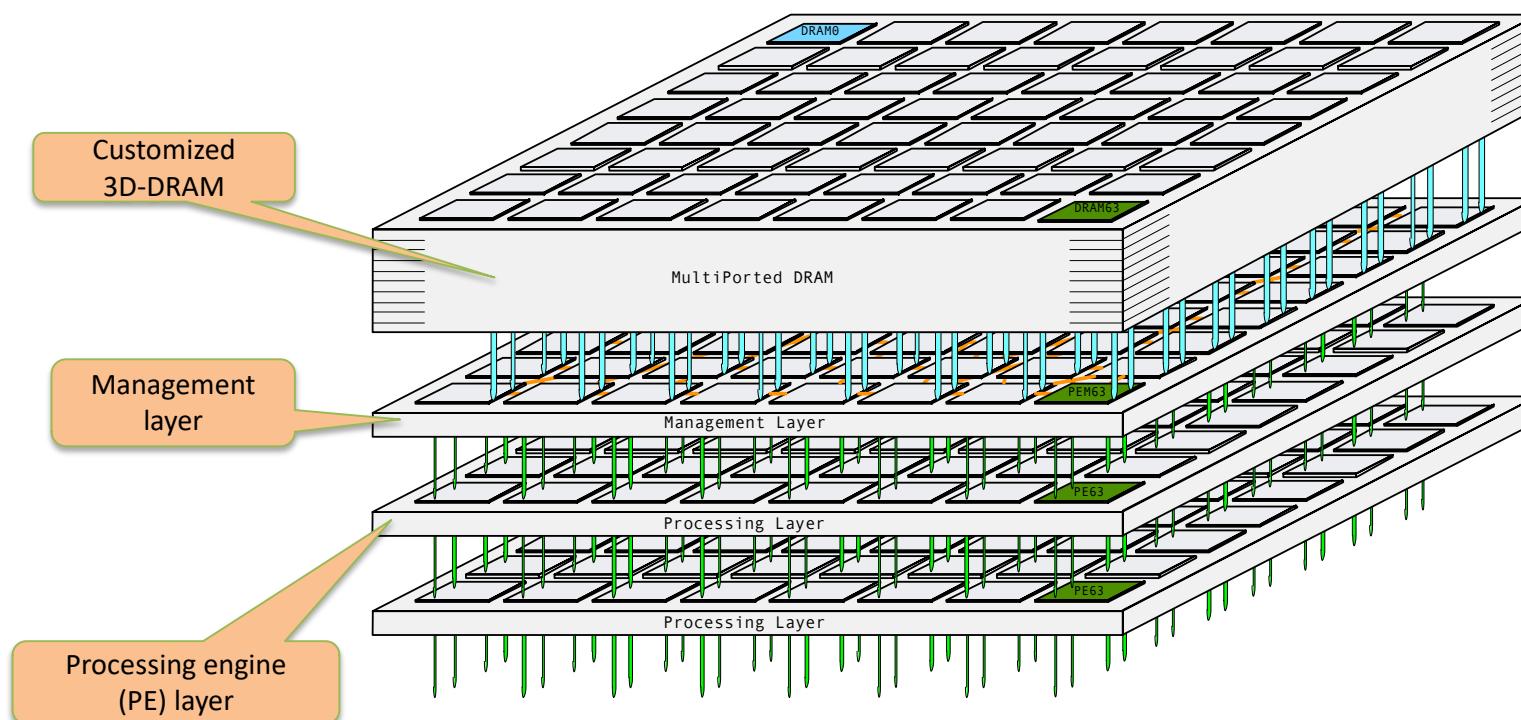
System Solution

- The 3D-DRAM is the Tezzaron DiRAM4
 - 64 disjoint 1Gb memory ports
 - System has 64 sub-systems each operating on one memory port
- Suggested Customizations for a 3D-DRAM
 - Standard DiRAM4 port is 32bits @ 1GHz
 - We suggest widen to 2048 bits and use high-density TSVs
 - Entire page in one access using burst-of-2
 - Raw bandwidth ~2Tbps per port

System Solution

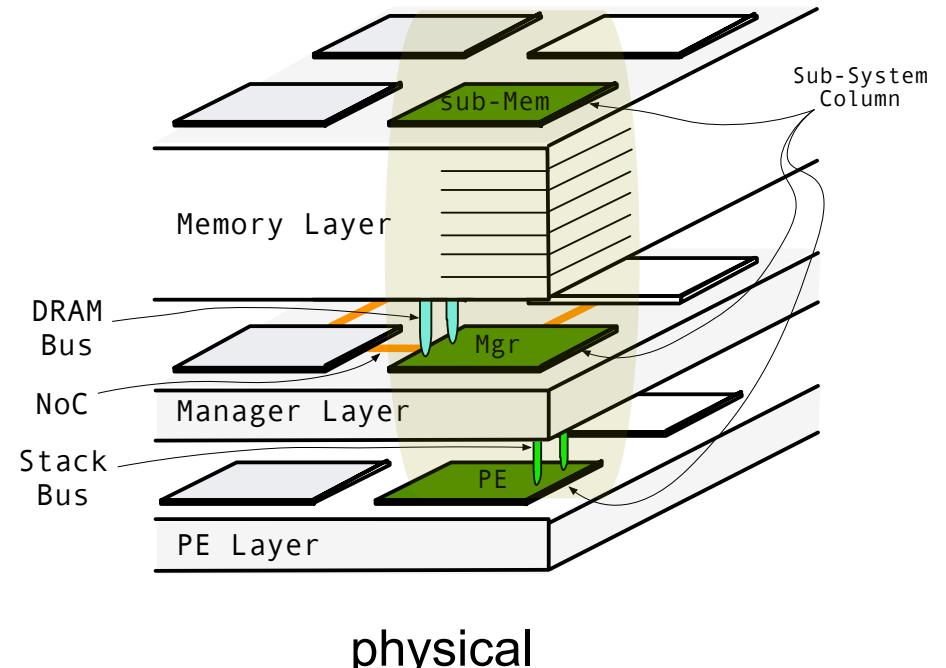
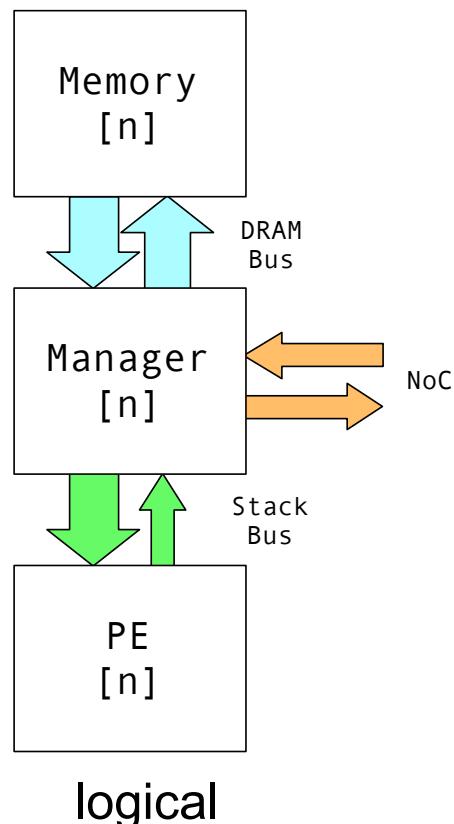
- System has 64 sub-systems (SSC) each operating on one memory port
 - ANNs split over sub-systems
 - SSCs communicate over local NoC
- Sub-system (SSC) has a Manager and PE
 - Instruction Decode in Manager
 - contains information on how to process a group of ANes
 - Wide Data-path between DRAM, Manager and PE
 - data for 32 execution lanes with two arguments per lane
 - able to perform a MAC operation between a weight and ANe state every cycle
 - Result path from PE to Manager
 - carries result(s) of processed ANes

3D Physical Configuration

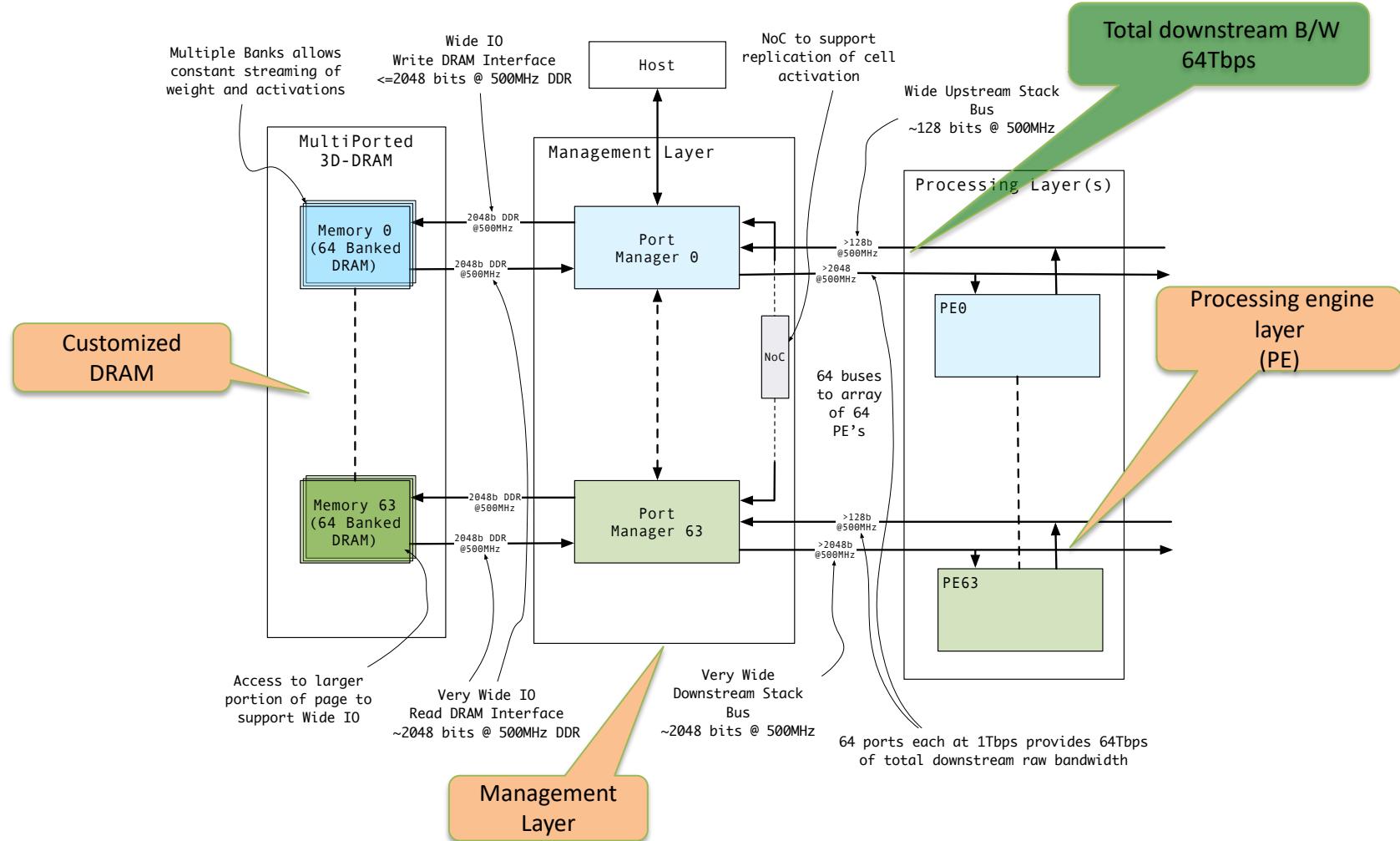


Sub-System Column (SSC)

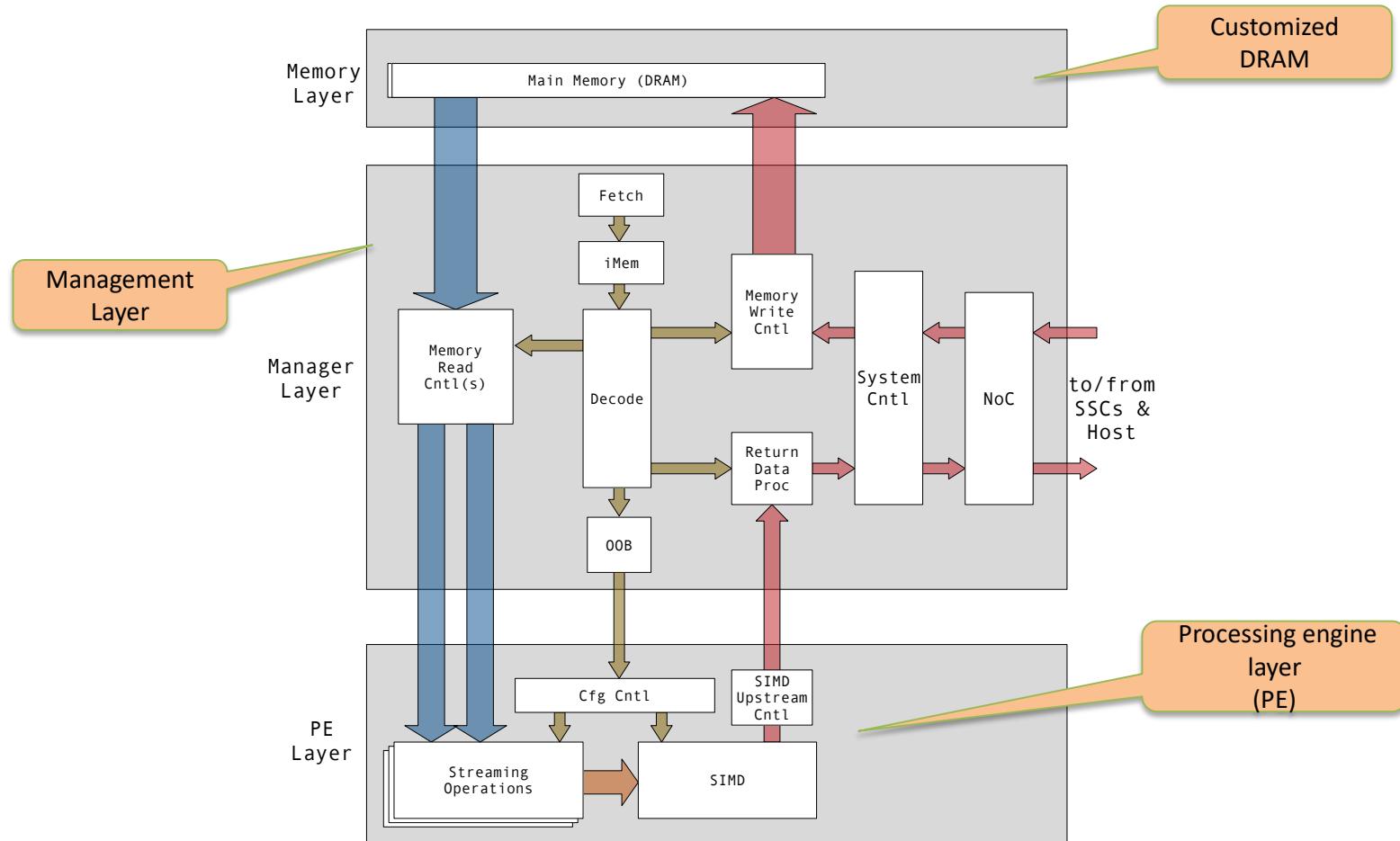
- Consists of DRAM, manager and processing engine (PE)



Solution Block Diagram



System Connectivity



Architecture Requirements

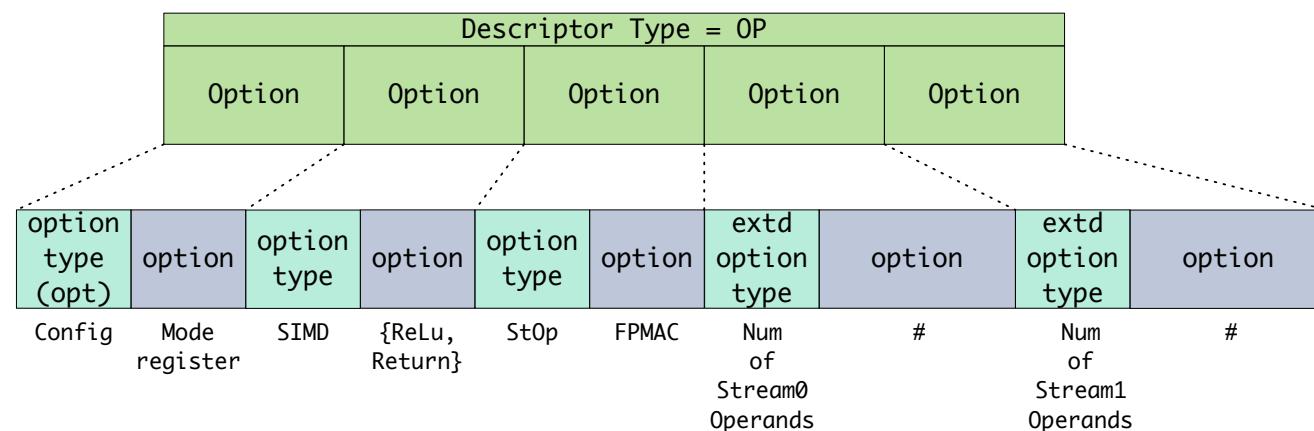
- Need to absorb DRAM latency
 - separate memory requests from read/write data
- Instruction spawns multiple commands to dependent tasks
 - concurrently pre-fetch data, prepare PE, prepare Return Data Processor
 - all operations have an associated tag
- Need to describe parameter and state storage
 - assume input stored in row-major fashion
- Network-on-Chip
 - needed for ANe state replication to all dependent SSCs

Instructions

- Instructions consist of descriptors

Operation Descriptor	arg0 Read Descriptor	arg1 Read Descriptor	Result Write Descriptor
----------------------	----------------------	----------------------	-------------------------

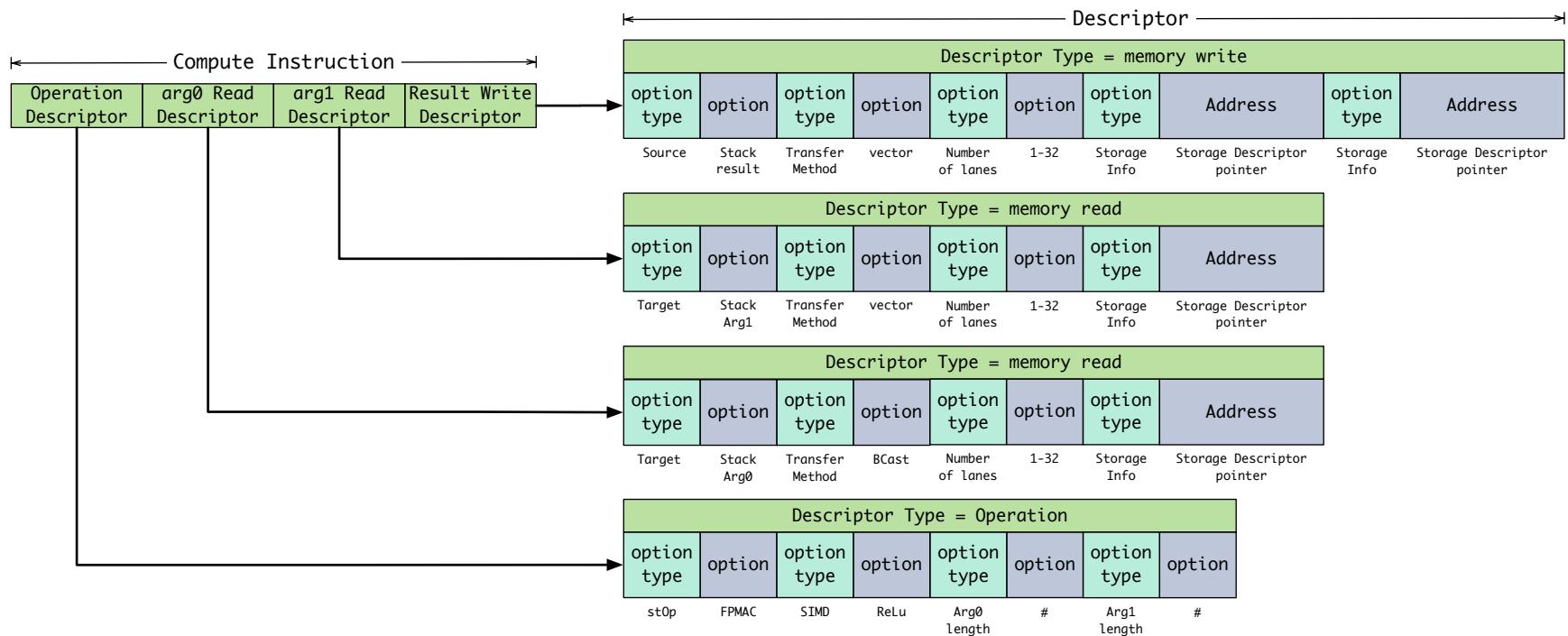
- Descriptors used to control modules which take part in the instruction
- Descriptors contain tuples which contain the details



Instruction Decode

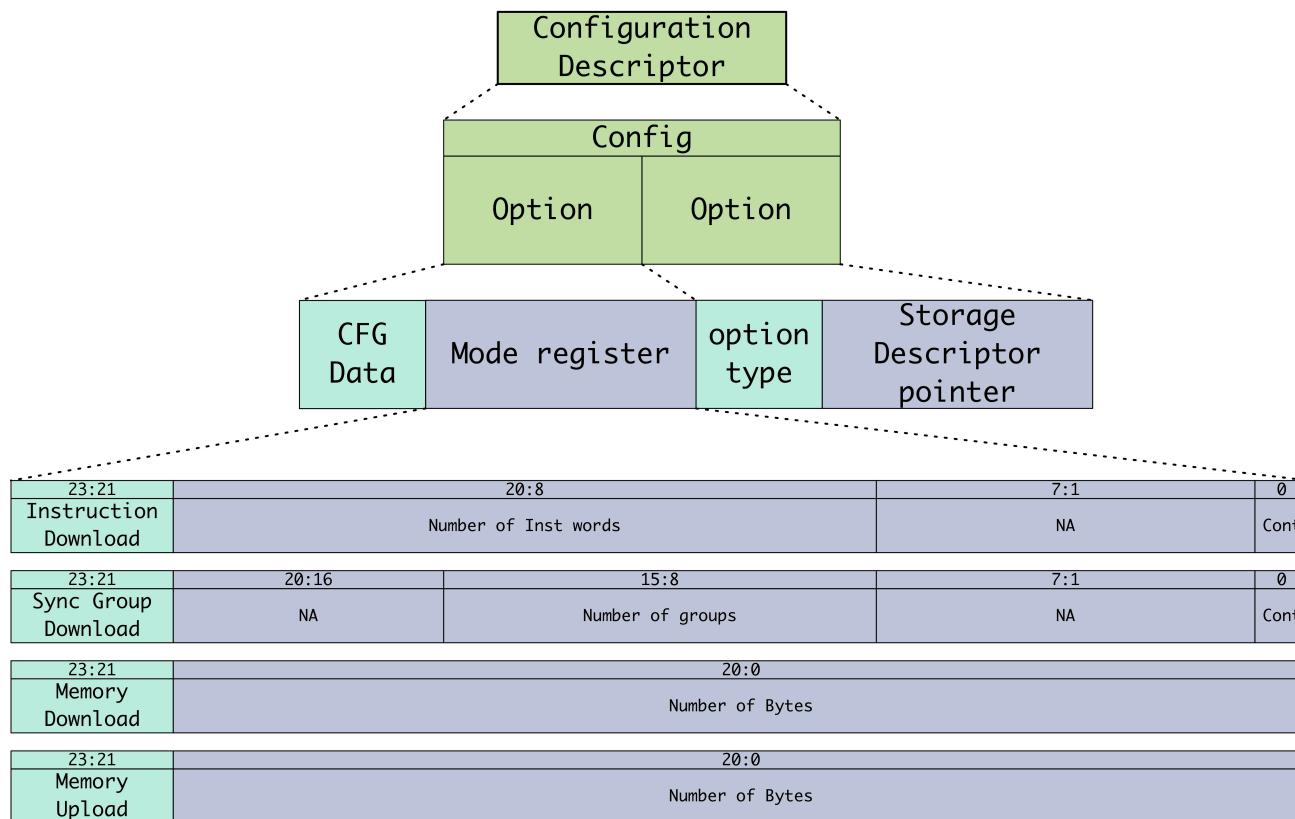
- Two types of instruction
 - Configuration
 - Compute
- Manager sends descriptors to dependent block(s)
 - may not parse descriptor if not relevant to manager
 - manager knows descriptor dependencies and sends to modules participating in instruction
- Configuration Instruction
 - send descriptor to manager control block
 - used for synchronization and data transfers
- Compute Operation
 - send descriptor(s) to blocks participating in ANe state computation
 - OP
 - send descriptor to PE
 - Memory Read
 - identify target execution lane and send descriptor to one of two memory read controller
 - Memory Write
 - send descriptor to return data processor

Compute Instruction



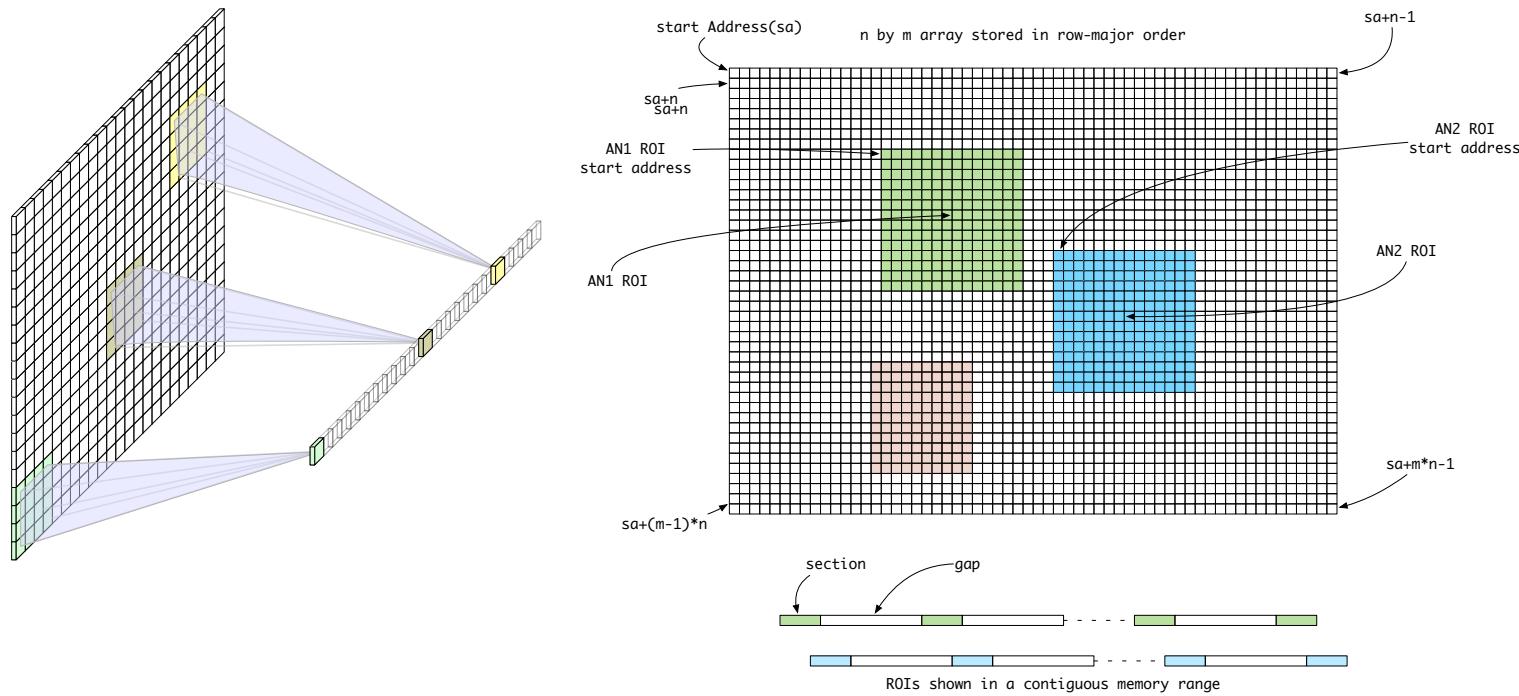
Configuration

- Supports both synchronization and data transfer



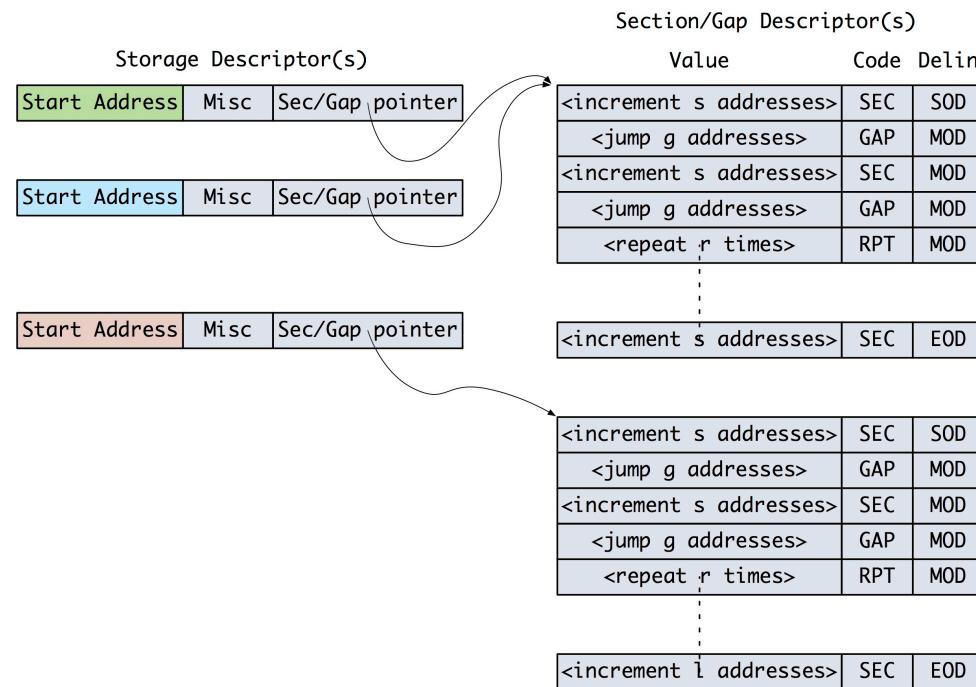
Data Storage and access

- Inputs and ANe states stored in row-major
 - two address increment mechanisms
- Weights for a group of ANes are interleaved



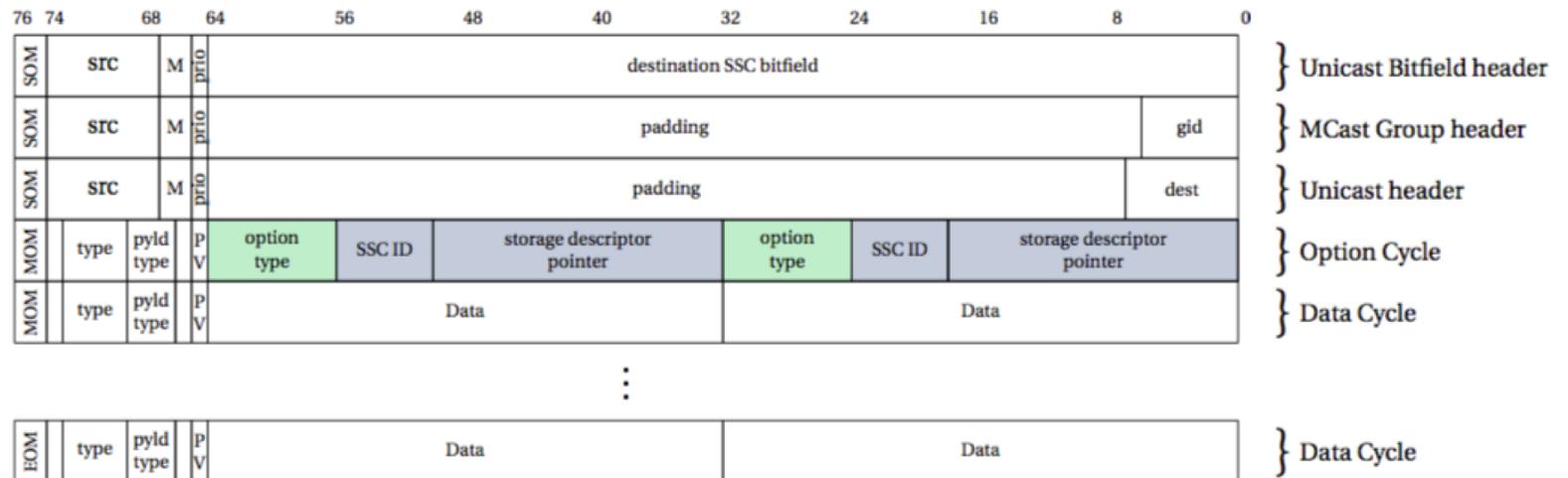
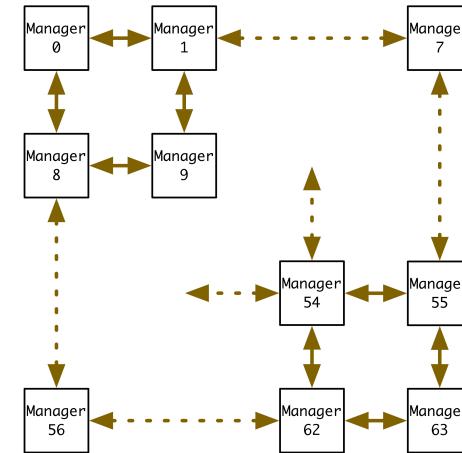
Storage descriptors

- Used to describe how data is read/written
- Used for both parameter and ANe state storage
 - instruction contains pointer to storage descriptor
 - each SSC has a set of storage descriptors
 - instruction/NoC carry pointer, not address, length etc.



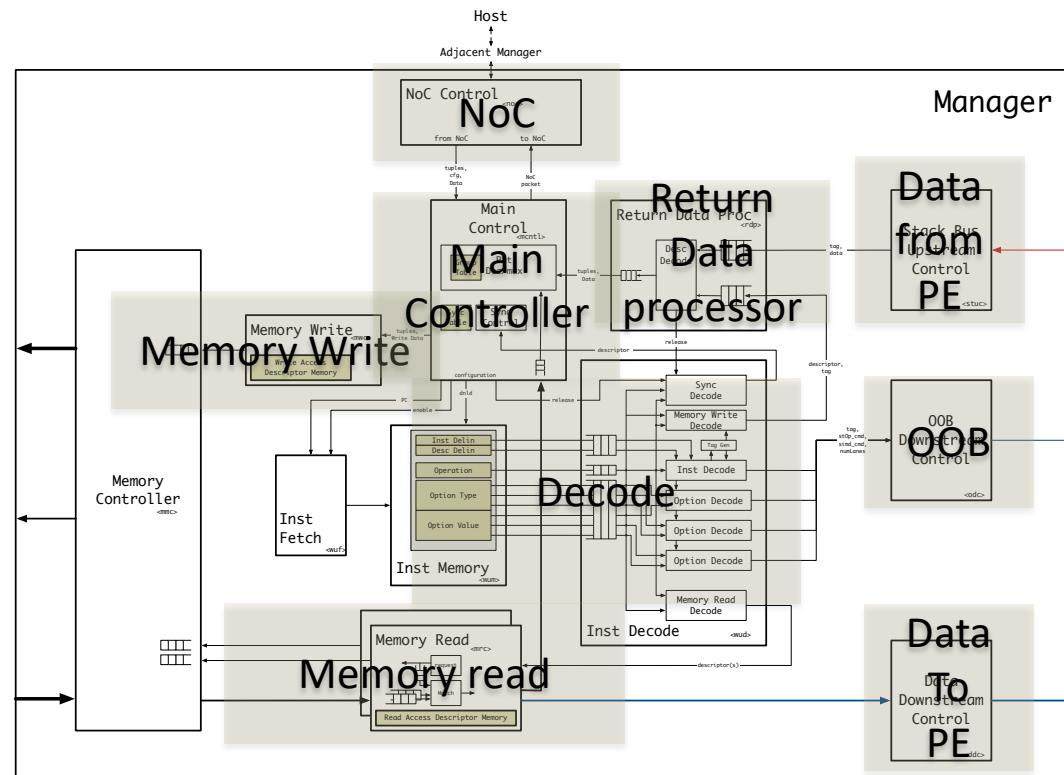
NoC

- Four ports
 - connections to adjacent SSC
 - fixed routing table
 - can be reconfigured based on performance
 - good area for additional research



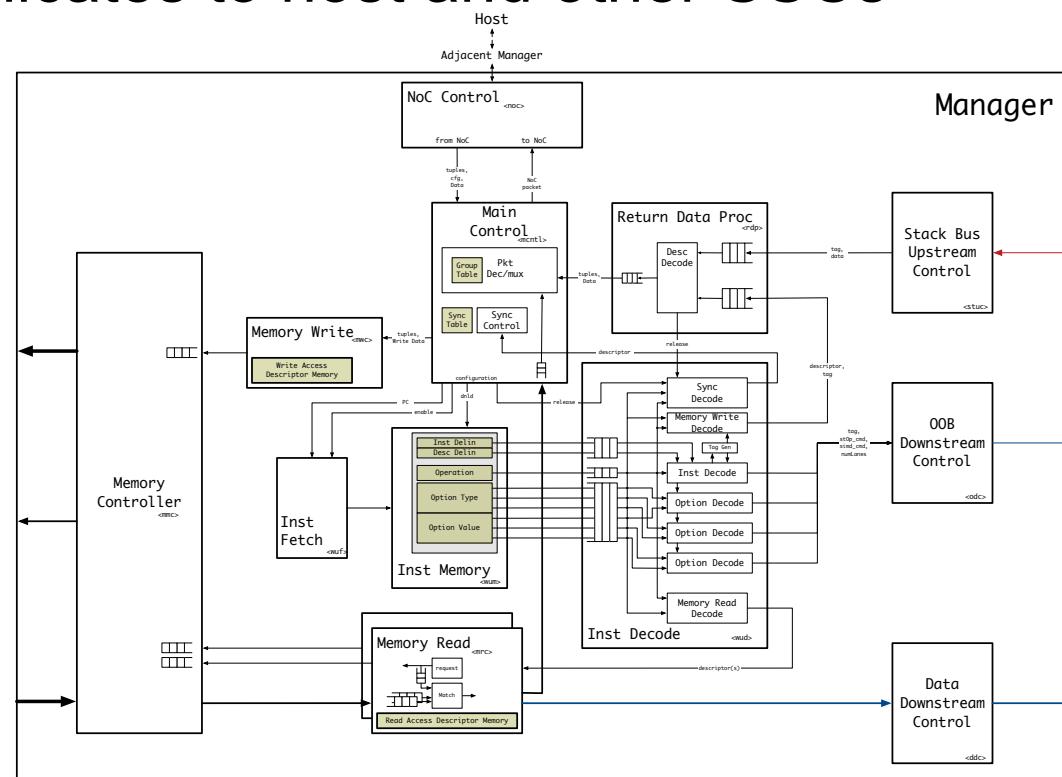
Manager Layer

- Decodes instructions
 - sub-descriptors are sent to dependent blocks
- Reads and writes to main memory
- Communicates to host and other SSCs



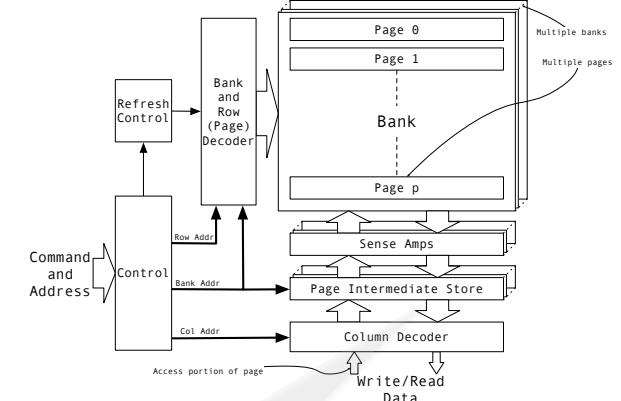
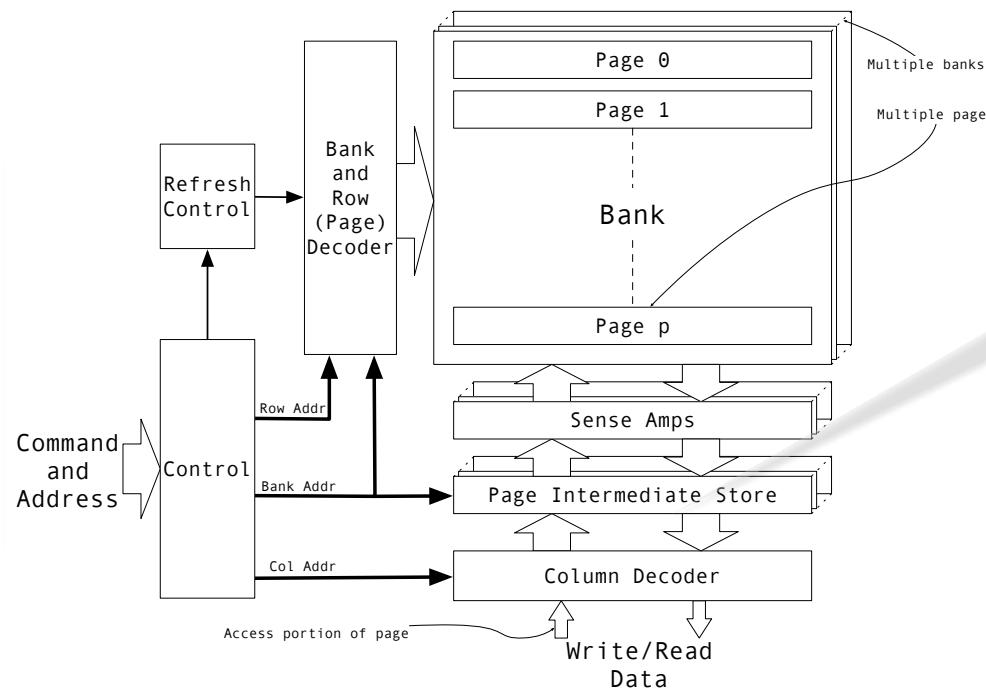
Manager Layer

- Decodes instructions
 - sub-descriptors are sent to dependent blocks
- Reads and writes to main memory
- Communicates to host and other SSCs



DRAM Customizations

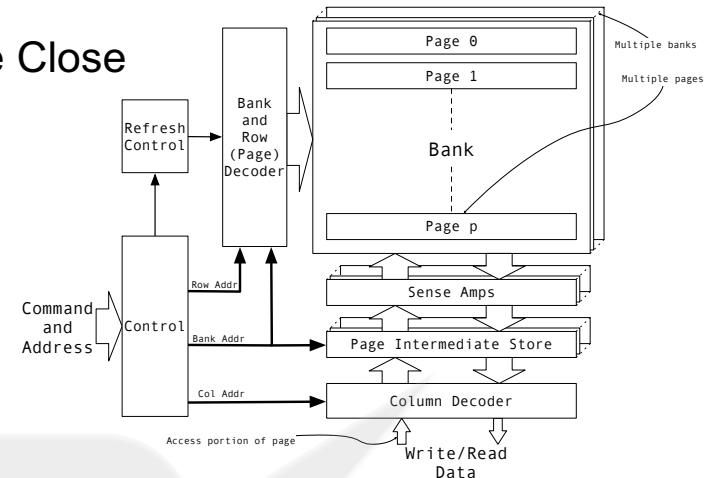
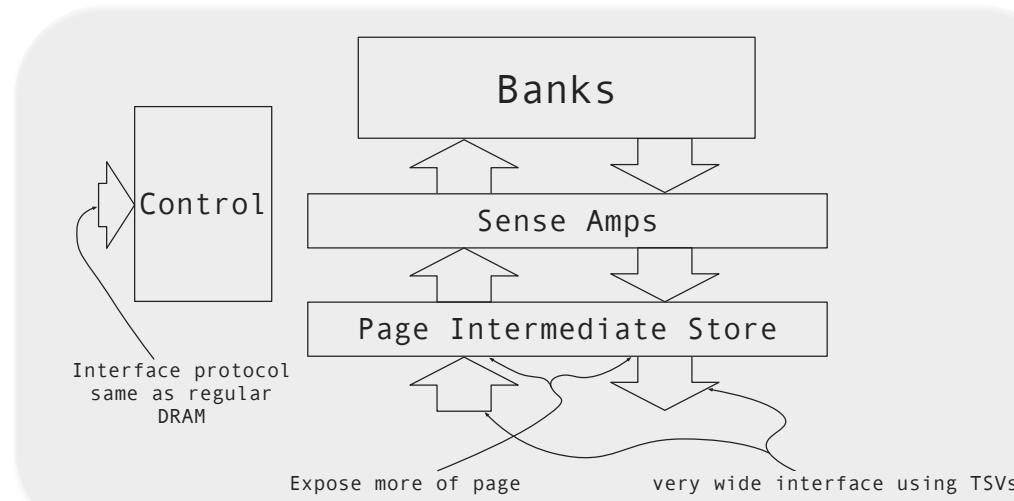
- Expose more of the page
 - Each burst-of-2 read transfers entire page
 - Need to accommodate idle time during Page Close
 - Employ TSVs for wide bus
- Provide write masks
 - To avoid Read-Modify-Write



Customize
Access

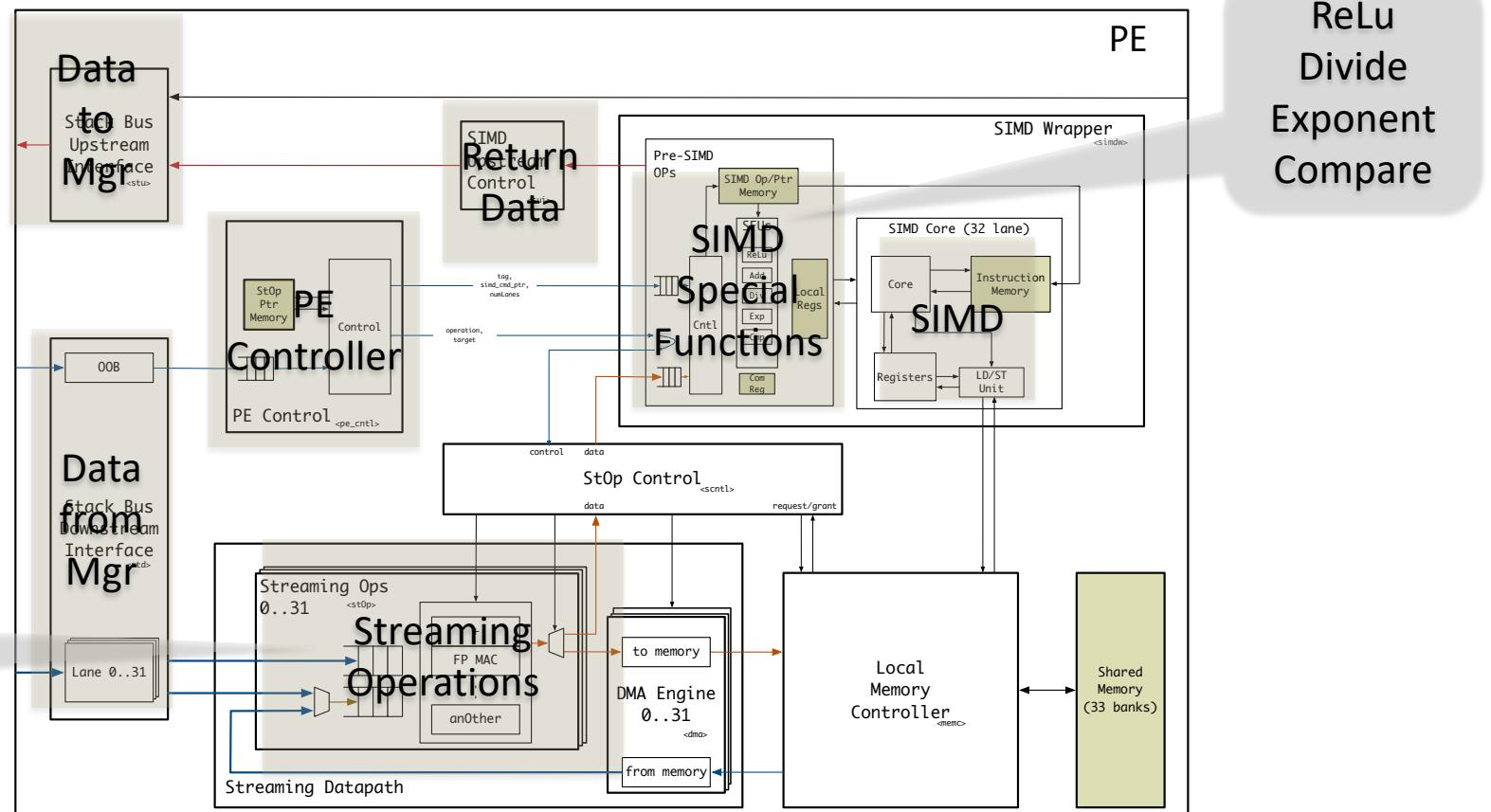
DRAM Customizations

- Expose more of the page
 - Each burst-of-2 read transfers entire page
 - Need to accommodate idle time during Page Close
 - Employ TSVs for wide bus
- Provide write masks
 - To avoid Read-Modify-Write



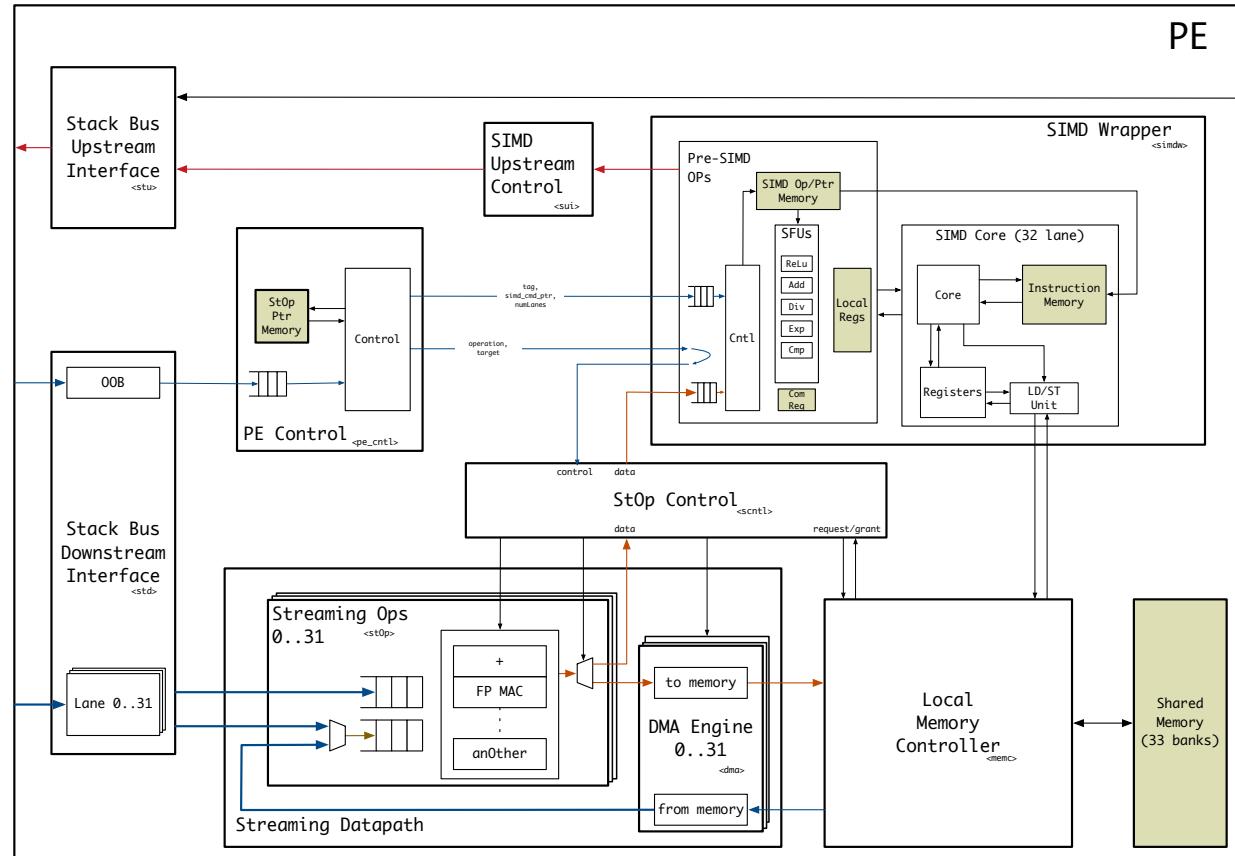
Processing (PE) Layer

- Streaming operations operate directly on data
- SIMD Wrapper performs post stOp tasks

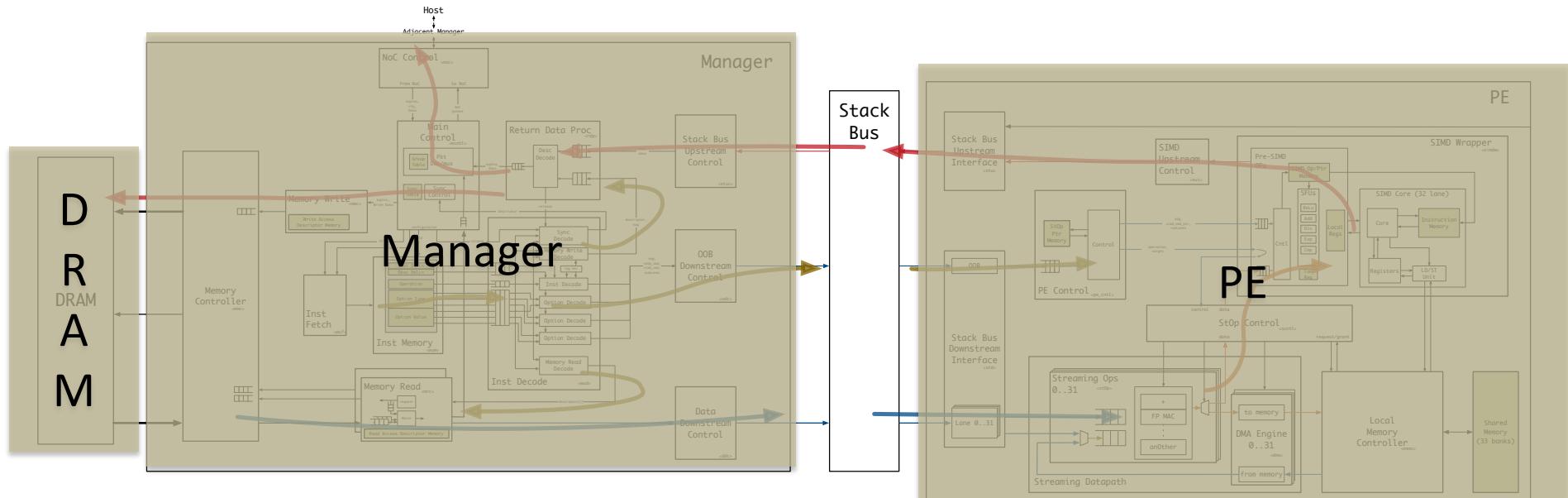


Processing (PE) Layer

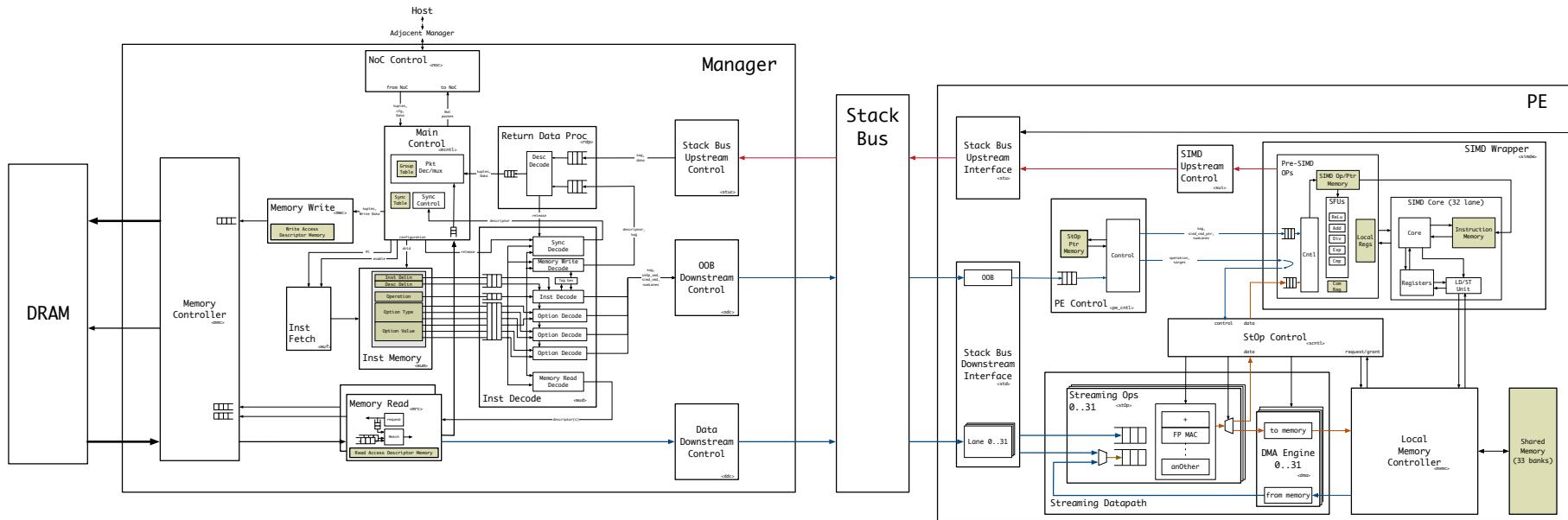
- Streaming operations operate directly on data
- SIMD Wrapper performs post stOp tasks



Detailed Block Diagram



Detailed Block Diagram



Test Performance

- Tested against multiple fan-ins

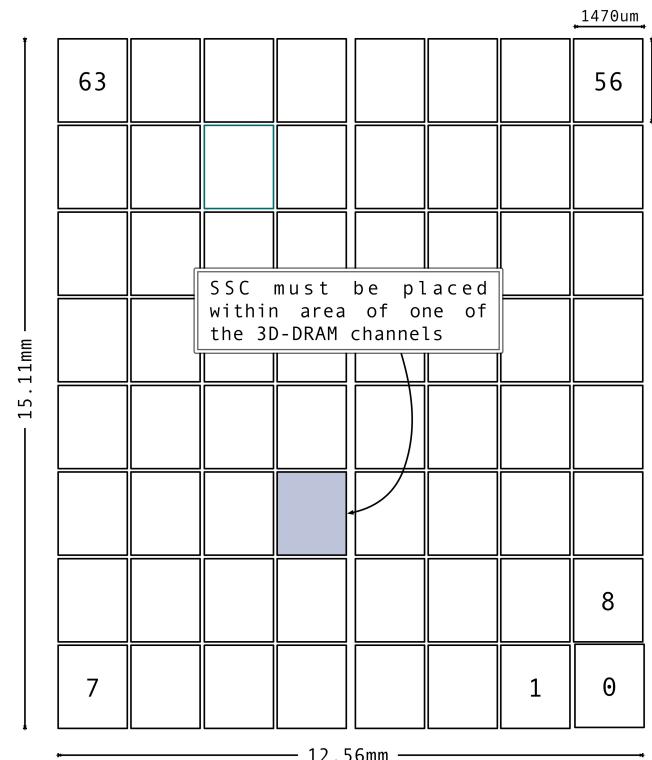
Test	System bandwidth (Tbps)
CONV2	22
CONV-294	23
CONV-300	25
CONV-500	26
CONV-1000	30
FC-350	26
FC-500	27
FC-1000	30
FC-7	31

Baseline ANN Performance

Layer	fanin	Equivalent test	Percentage of Operations	Expected bandwidth
1	363	CONV-300	44%	10.9
2	4	previous layer		
3	2400	CONV-1000	28%	8.5
4	4	previous layer		
5	2304	CONV-1000	10%	3
6	3456	CONV-1000	10%	3
7	3456	CONV-1000	7%	2
8	43264	FC-7	1%	0.2
9	4096	FC-7	1%	0.2
10	4096	FC-7	<1%	0.05
		Total	>27.7	

AREA

- Area based on DiRAM4
 - SSC needs to fit into footprint of DiRAM4 sub-memory
 - available area is 2.43 sq-mm



Power/Area Scaling assumptions

- Scaling based on synthesizing representative block at 28nm and 65nm
 - Designed using 65nm library, 28nm is target technology

- Area

Area Scaling 65nm -> 28nm	
Memory	2.79
Logic	2.68

- Power

Scaling			
65nm to 28nm			
	Internal	Net switch	Leakage
Logic	5.07	1.21	4.12E-04
memory	5.07		

Manager AREA

- **Blocks**

	Manager		
	Distribution	65nm (um)	28nm(um)
Memory Controller	20.6%	994000	363478
NoC	6.9%	335000	122500
Read Control	47.1%	2270176	830141
Write Control	6.7%	321000	117381
Instruction Proc	1.7%	83570	30559
Read data proc	1.6%	78000	28522
System Controller	1.6%	77000	28157
TSV	6.9%	333750	333750
Misc	6.8%	330000	120672
	100.0%	4822496	1975160

- **Utilization**

	Manager	
	65nm	28nm
Area used	4822496	1975160
Area available	6661496	2434320
Utilization	72%	81%
Utilization w/o TSV	71%	78%

PE AREA

- **Blocks**

	Processing Engine		
	Distribution	65nm	28nm
Operation Decode	3.1%	105000	38370
Return Data Control	1.5%	49000	17906
SIMD Wrapper	15.1%	507000	185274
SIMD	17.9%	600000	219259
Streaming Ops	40.0%	1344000	491140
StOp Control	1.9%	64000	23388
LM	16.4%	550000	200987
TSV	3.7%	124750	124750
Misc	0.4%	15000	5481
Total	100.0%	3358750	1306555

- **Utilization**

	Processing Engine	
	65nm	28nm
Area used	3358750	1306555
Area available	6661496	2434320
Utilization	50%	54%
Utilization w/o TSV	49%	51%

Area placement study

- Placed and routed without DRC or LVS
- Routing congestion showed minimal hotspots
- Parasitics used in Primetime for power analysis



Power Summary

- Power estimates based on CONV-294 testcase simulation
 - back to back operations
 - accumulate DRAM access and use DiRAM4 datasheet [xx]
 - TSV energy from [XX]
 - parasitics from layout
- designed using 65nm library
- 28nm is target technology
- power and area scaling based on synthesizing representative blocks

Parameters	
Frequency	500MHz
Test	CONV-294

Block	Power
Manager	42.55
PE	26.50
DRAM	4.51
DRAM TSVs	1.14
Stack Bus TSVs	0.74
Total	75.44

Comparison

- Best comparison is to NeuroStream and DaDianNao
 - Goldilocks ratio

	TPU	NeuroStream	DaDianao	GPU	this work
BW Utilization (Goldilocks ratio)	9028%	1270%	5%	451%	40%
	Coldest	Colder	Hot	Cold	Just right

- Scaled

	TPU	NeuroStream	DaDianao
Power at capacity Ratio	40	88	3549
	53%	117%	4732%
Power at bandwidth Ratio	3611	1117	167
	4815%	1490%	223%
Area at capacity Ratio	300	792	14889
	86%	226%	4254%

Summary

- Real world applications will require multiple artificial neural networks
 - current solutions consume significant power and real-estate
- A 3DIC solution that includes:
 - A proposed custom 3D-DRAM
 - Instructions and data structures
 - Purpose designed functions to accelerate a Artificial Neural Networks
 - Architecture to take advantage of 3D technology
- Overall performance and power improvement

Publication(s)

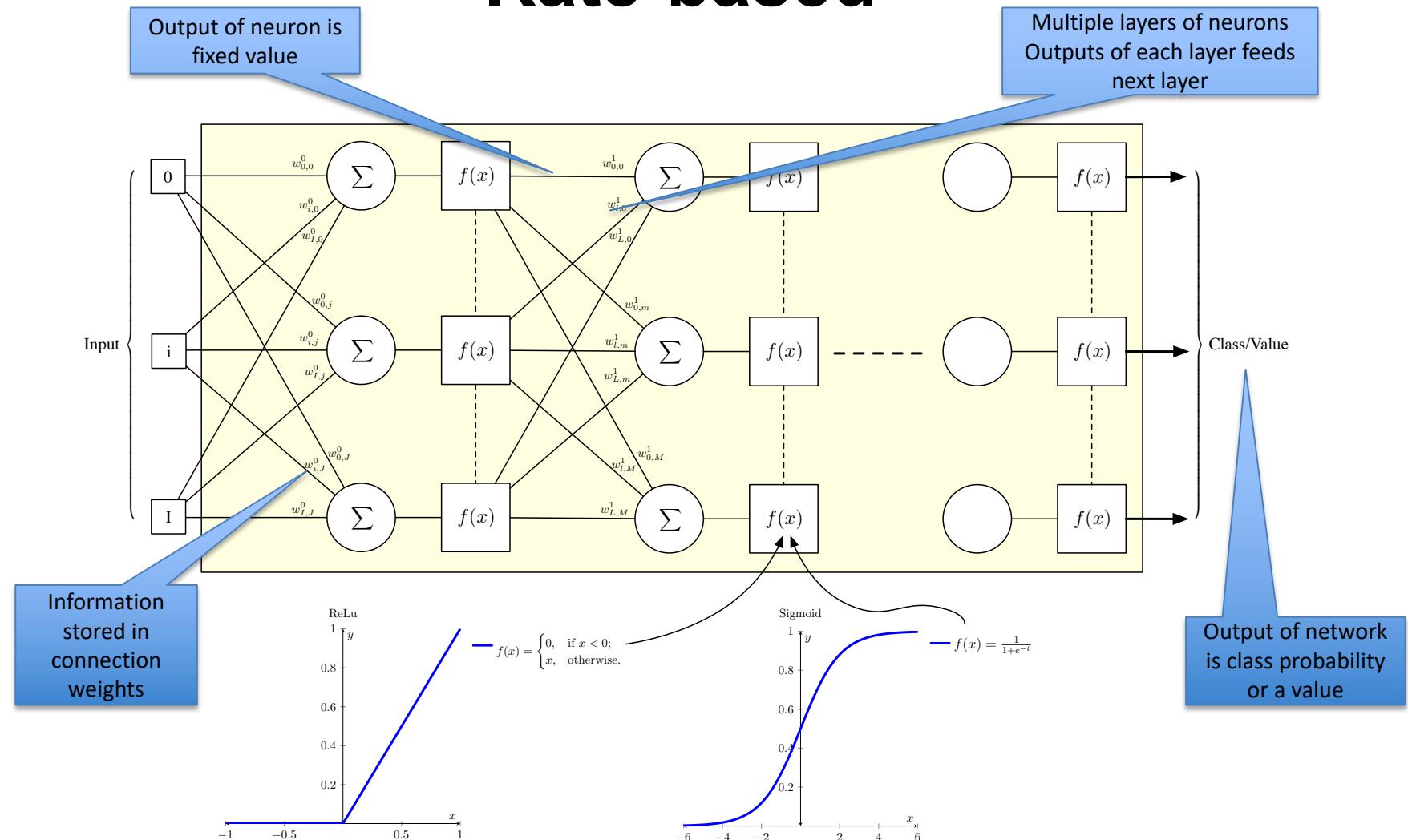
- Multi-ANN Edge System based on a Custom 3DIC DRAM - submitted to IEEE journal on emerging technologies and selected topics
- Using a 2-D Laser Scanner along with Cogent Confabulation as a Localized Navigation Aid
 - still trying to find an appropriate journal

References

Backup

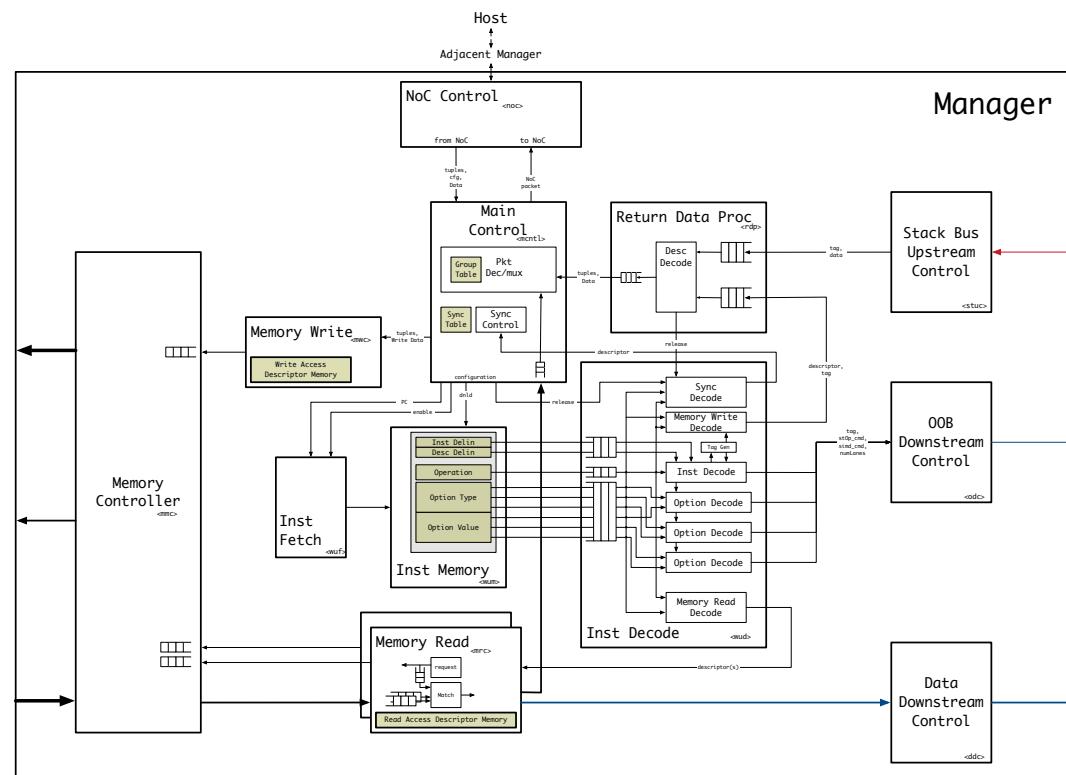
Artificial Neural Networks

Rate-based



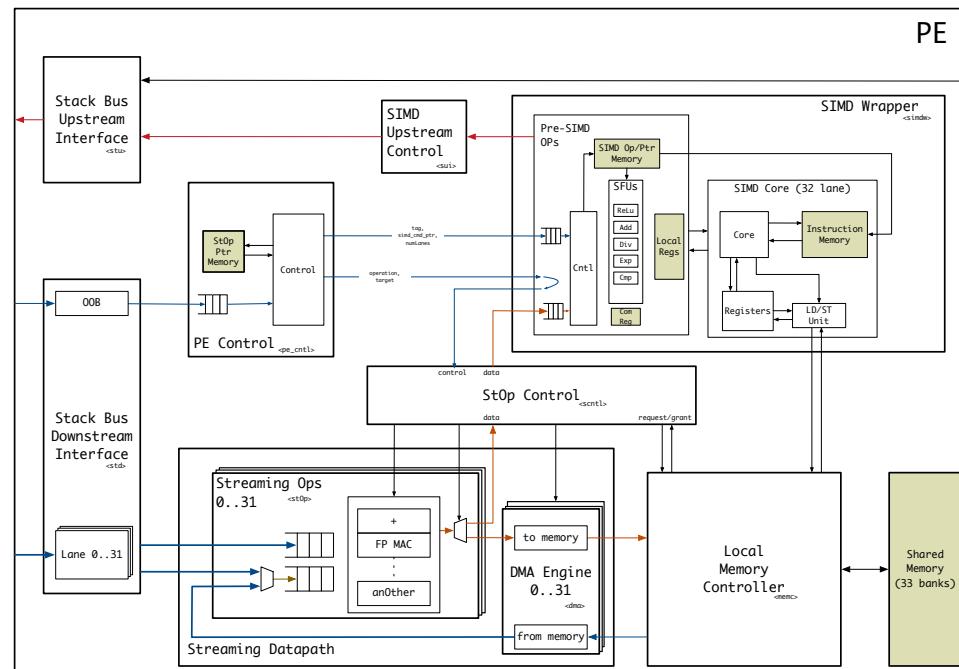
Manager Layer

- Decodes instructions
 - sub-descriptors are sent to dependent blocks
- Reads and writes to main memory
- Communicates to host and other SSCs



Processing (PE) Layer

- Streaming operations operate on data directly from the Stack bus
 - MAC, Multiply
- SIMD Wrapper performs post stOp tasks
 - add/divide/exponent, sends result back to manager

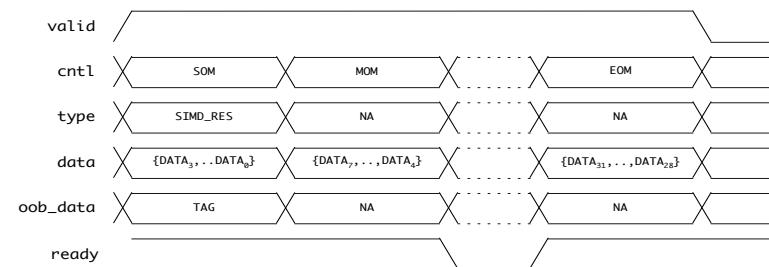
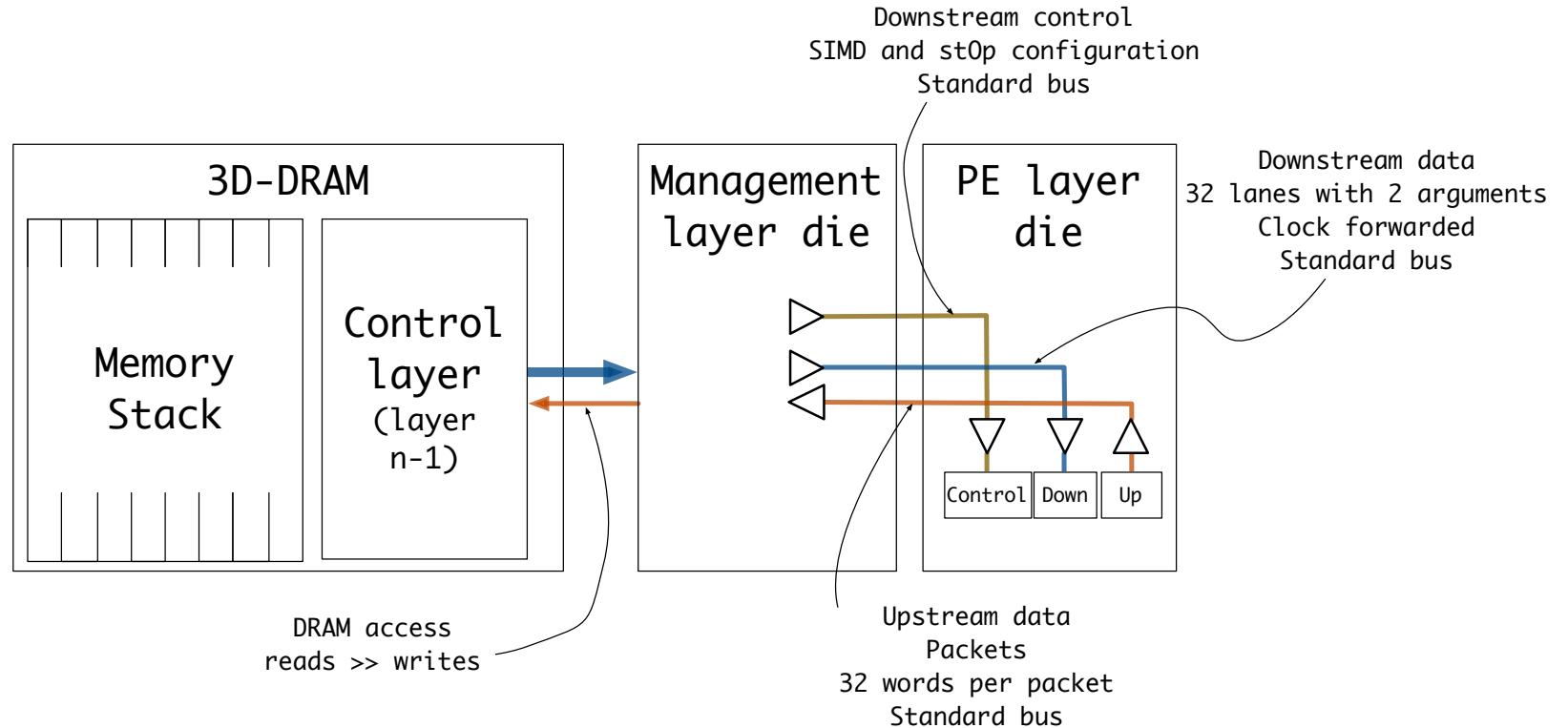


Power/Area Scaling assumptions

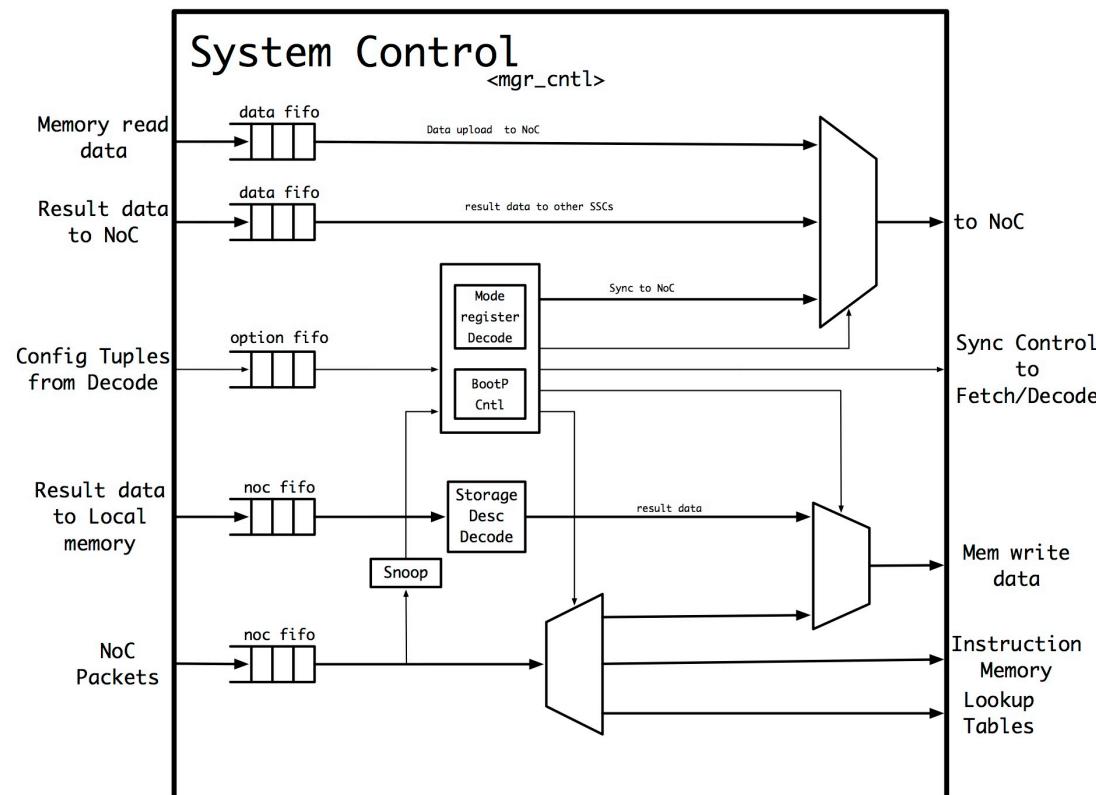
- Scaling based on synthesizing representative block at 28nm and 65nm
 - designed using 65nm library
 - 28nm is target technology

Synthesis numbers						
	65nm			28nm		
	Internal	Net switch	Leakage	Internal	Net switch	Leakage
Logic	66.9	1.53	2.02	13.2	1.26	4900
memory	2.36			0.0438		

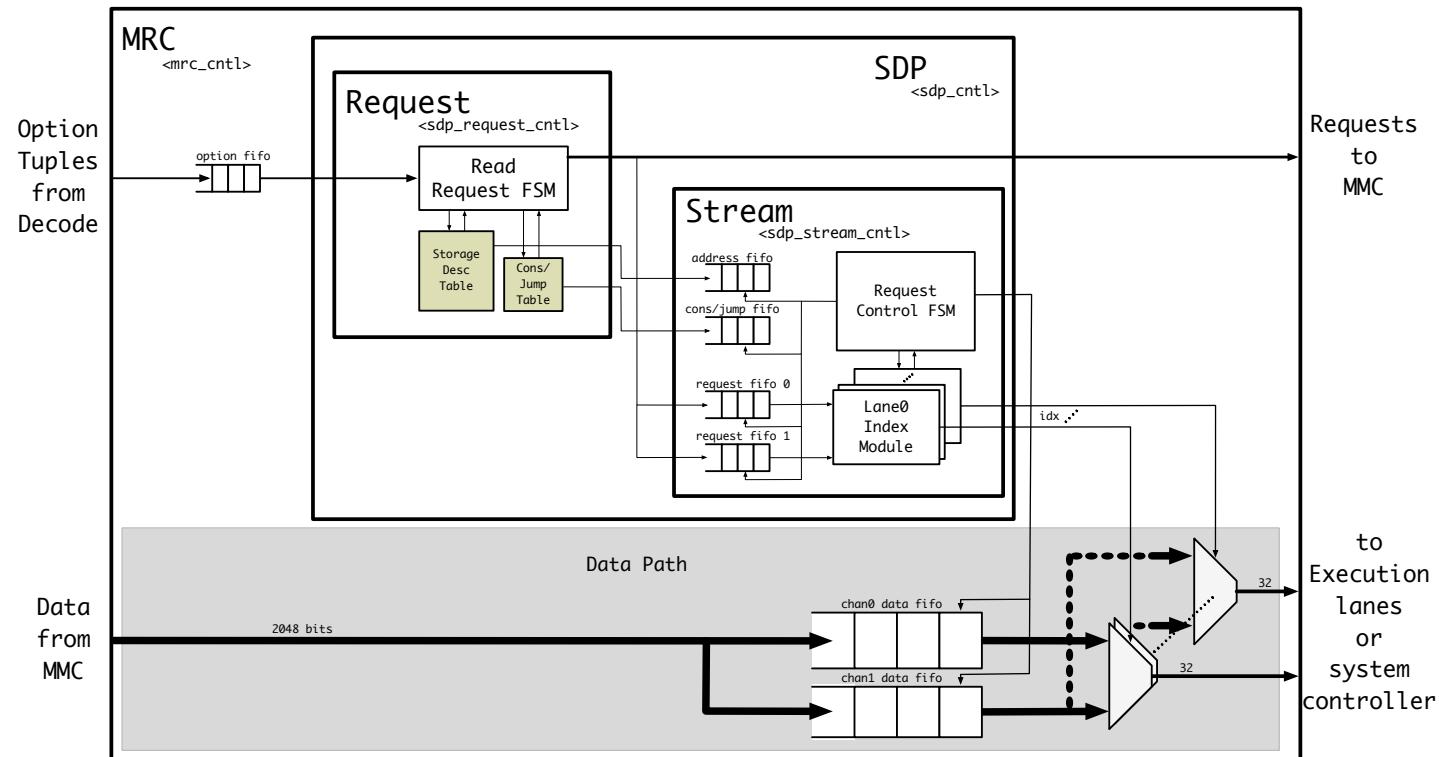
DRAM and Stack Buses



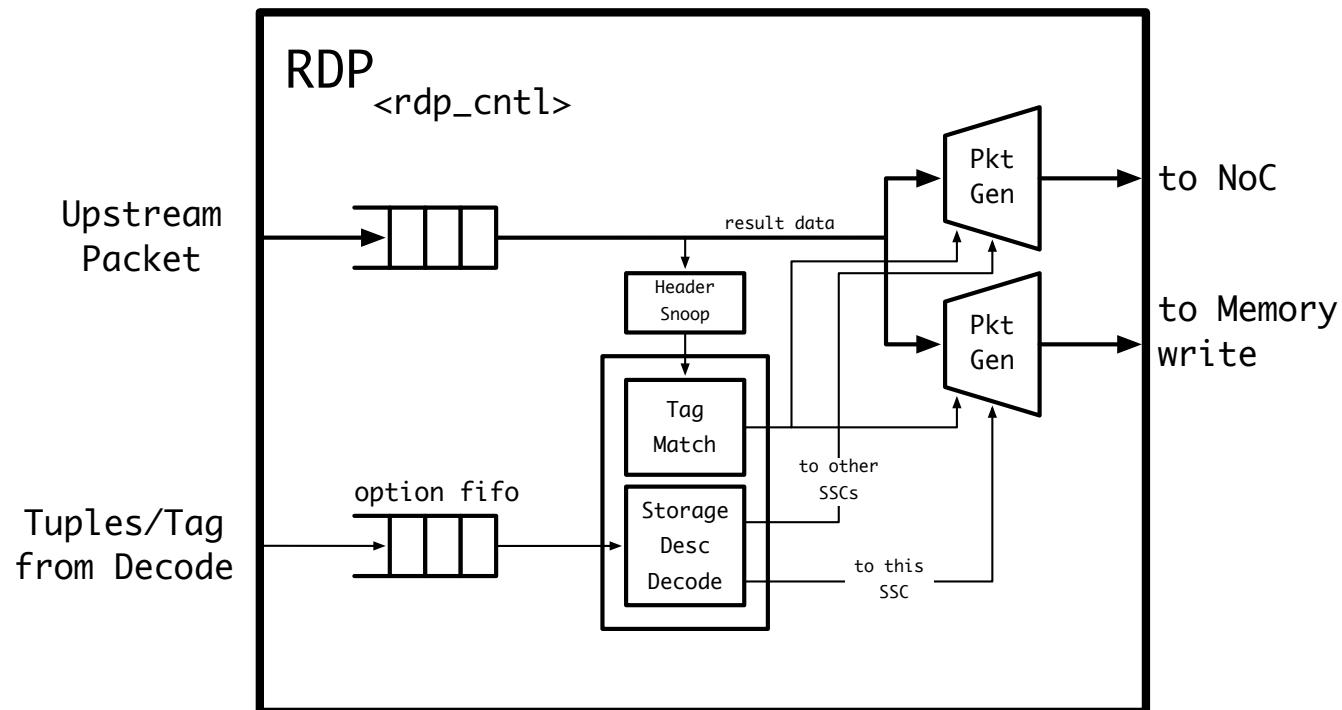
Manager Controller



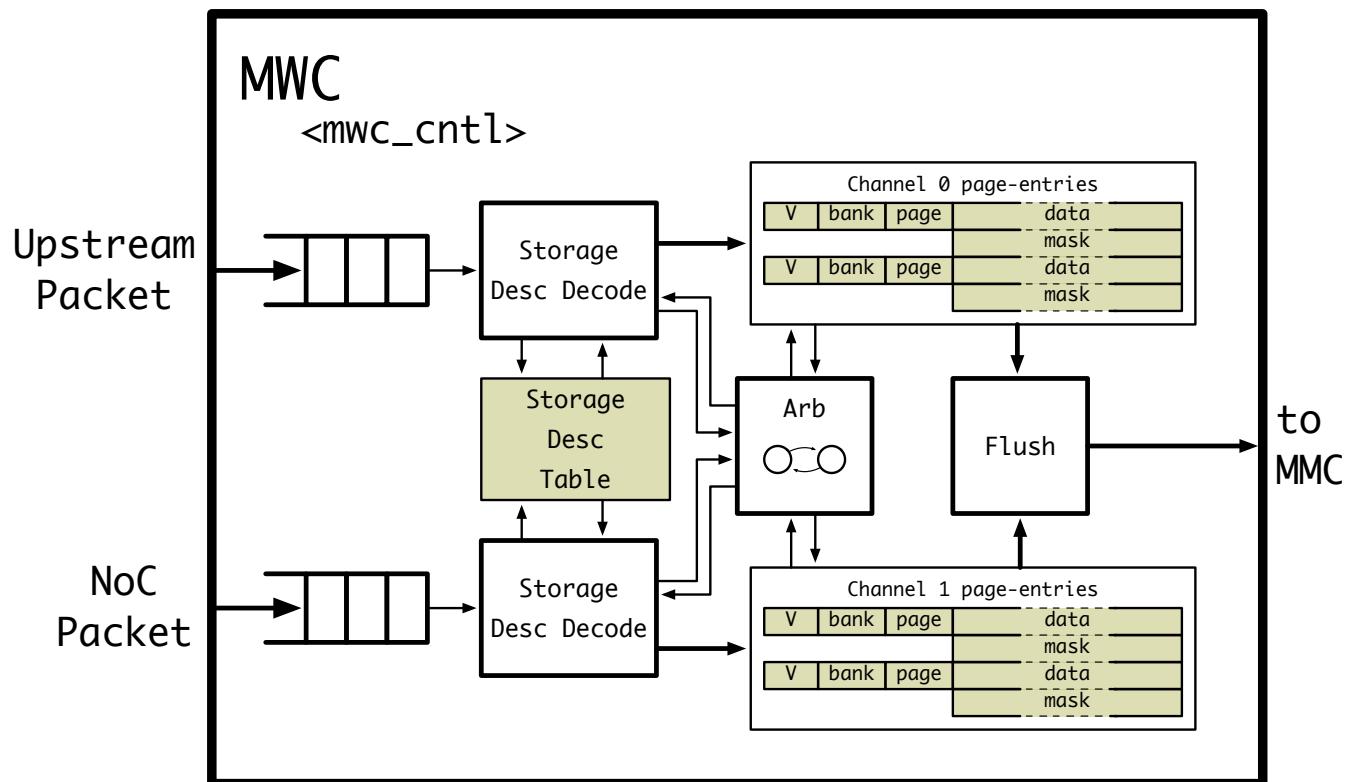
MRC



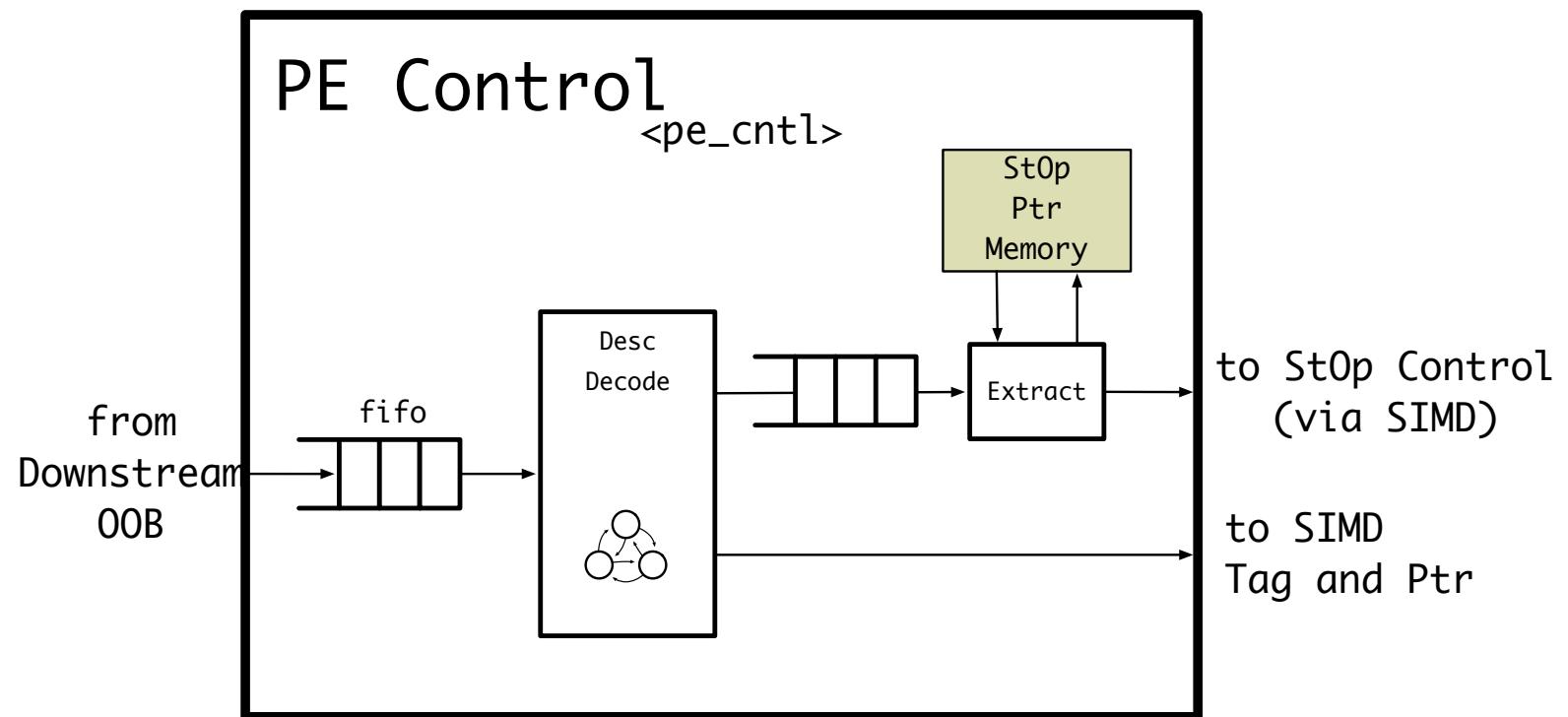
RDP



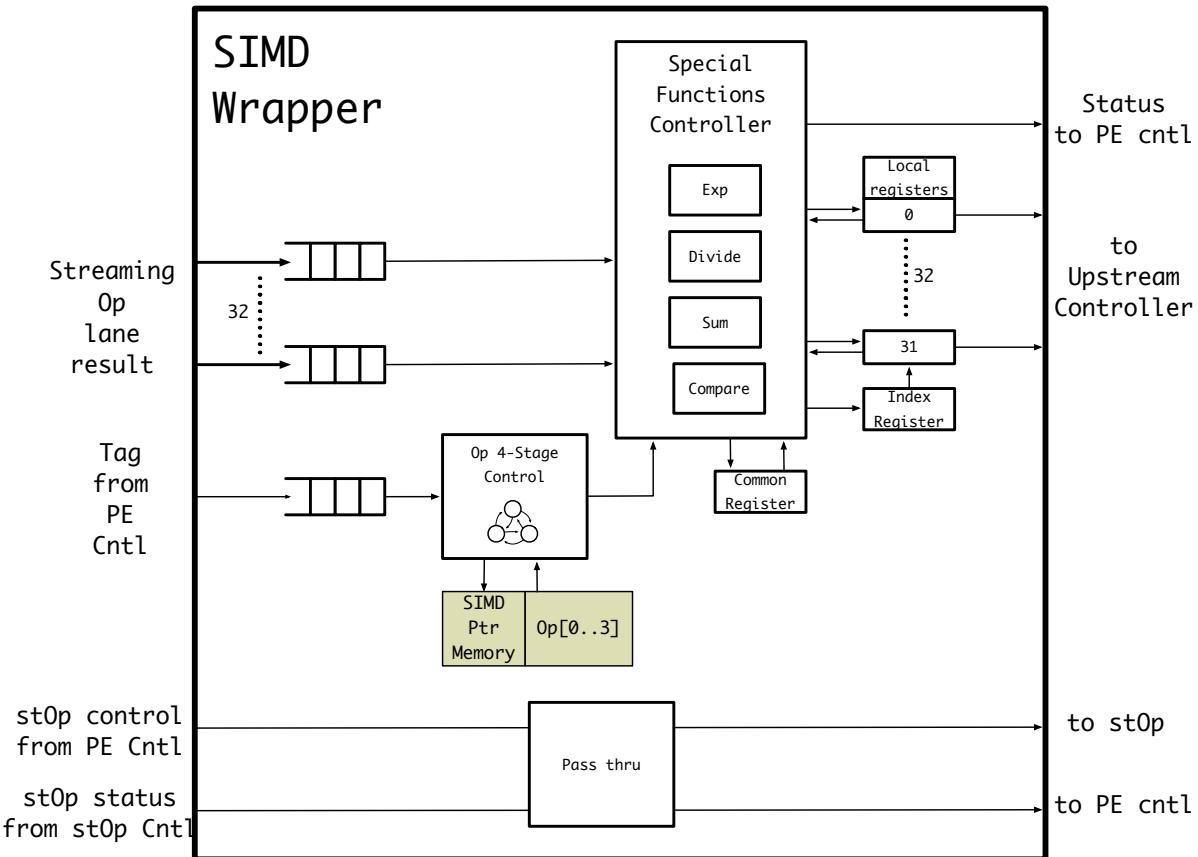
MWC



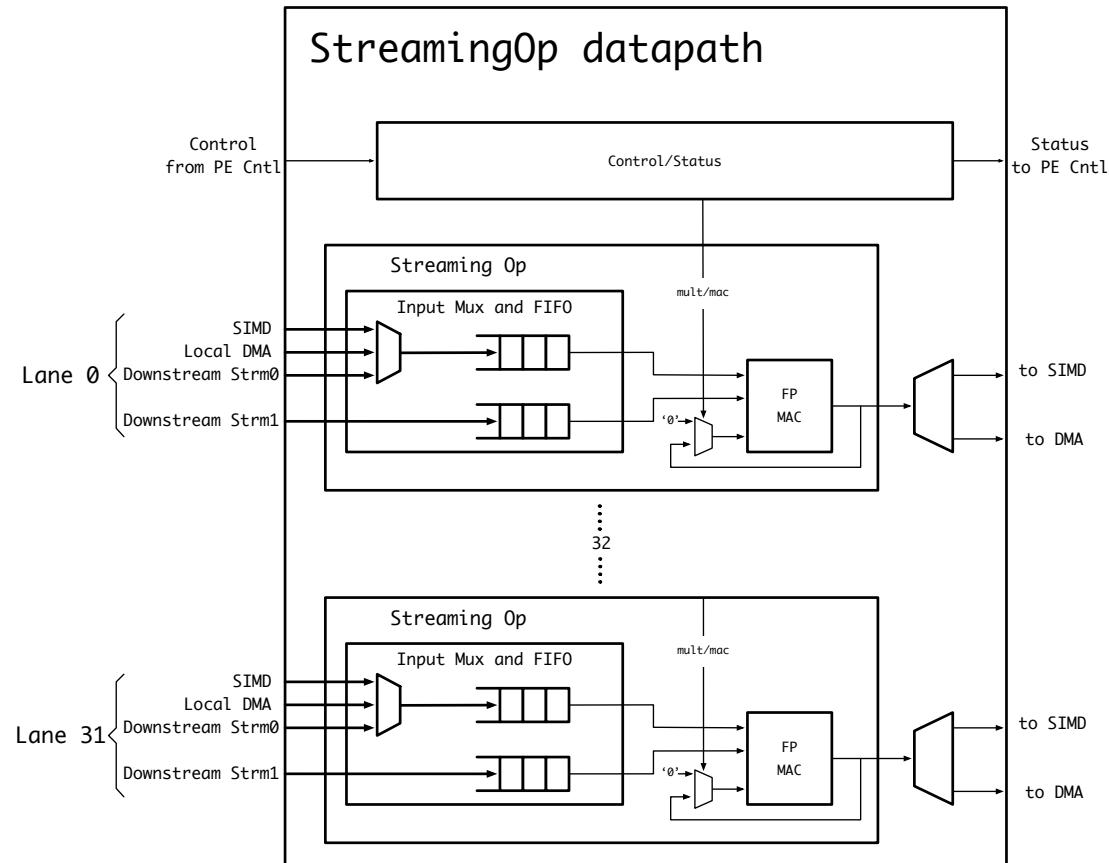
PE Controller



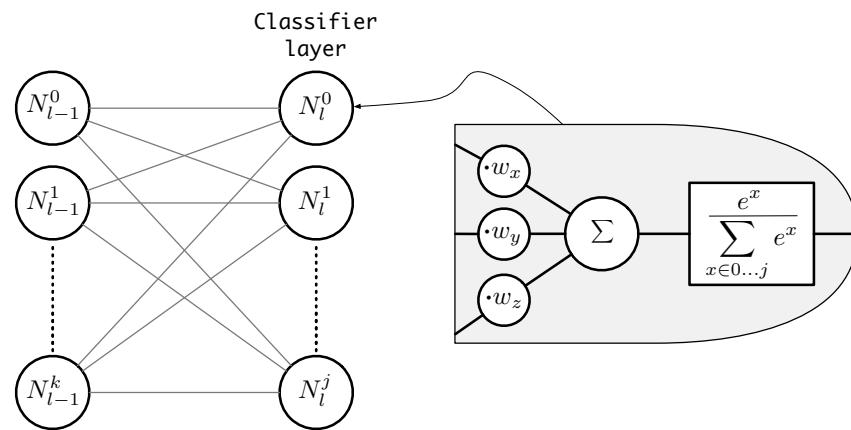
SIMD Wrapper



Streaming Operations

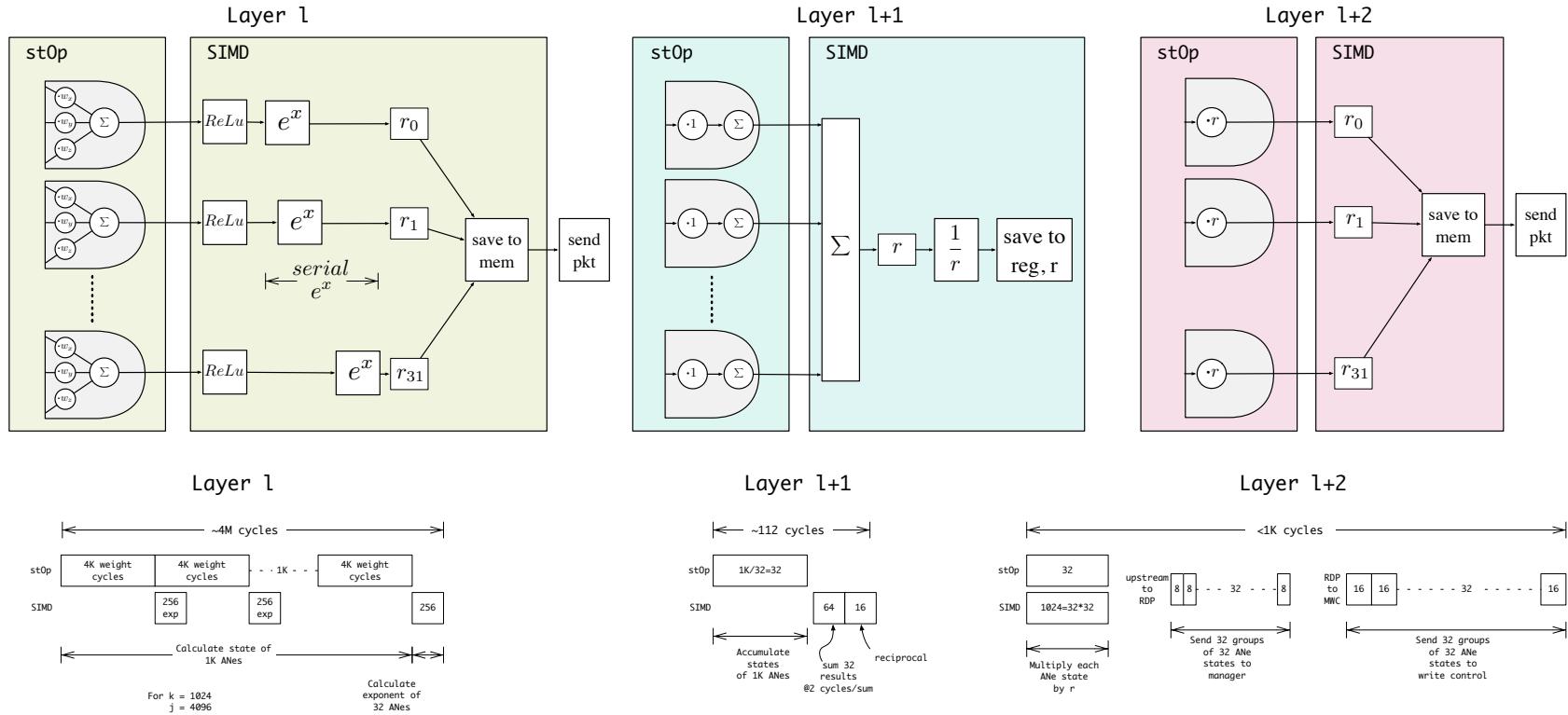


Classifier Implementation



- Separate into sub-layers

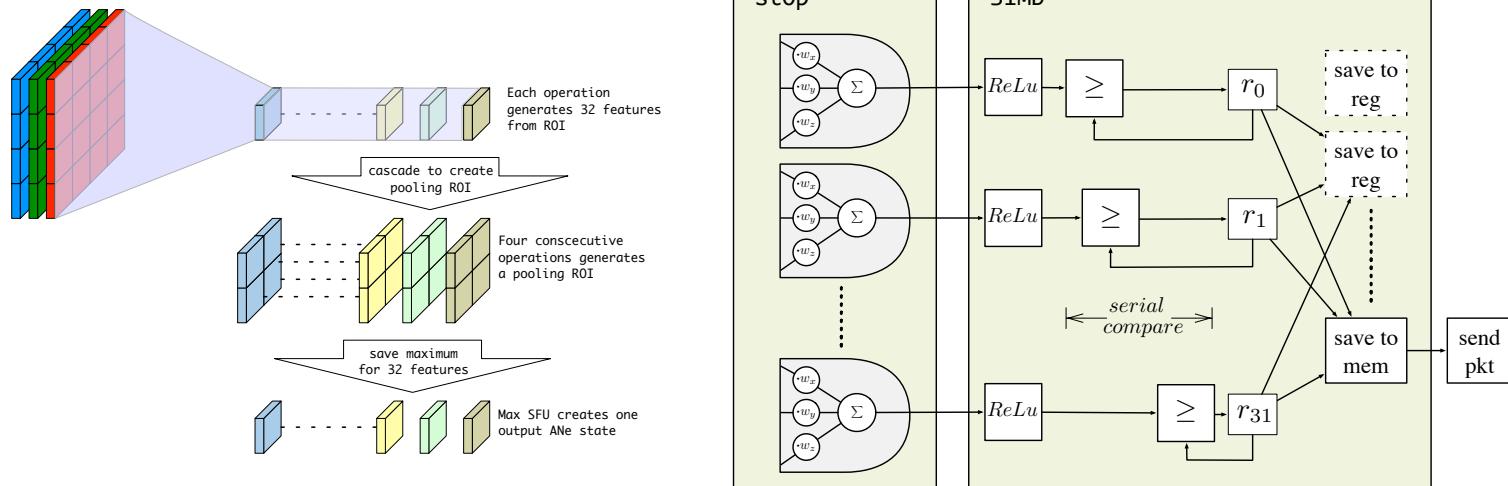
Classifier Implementation



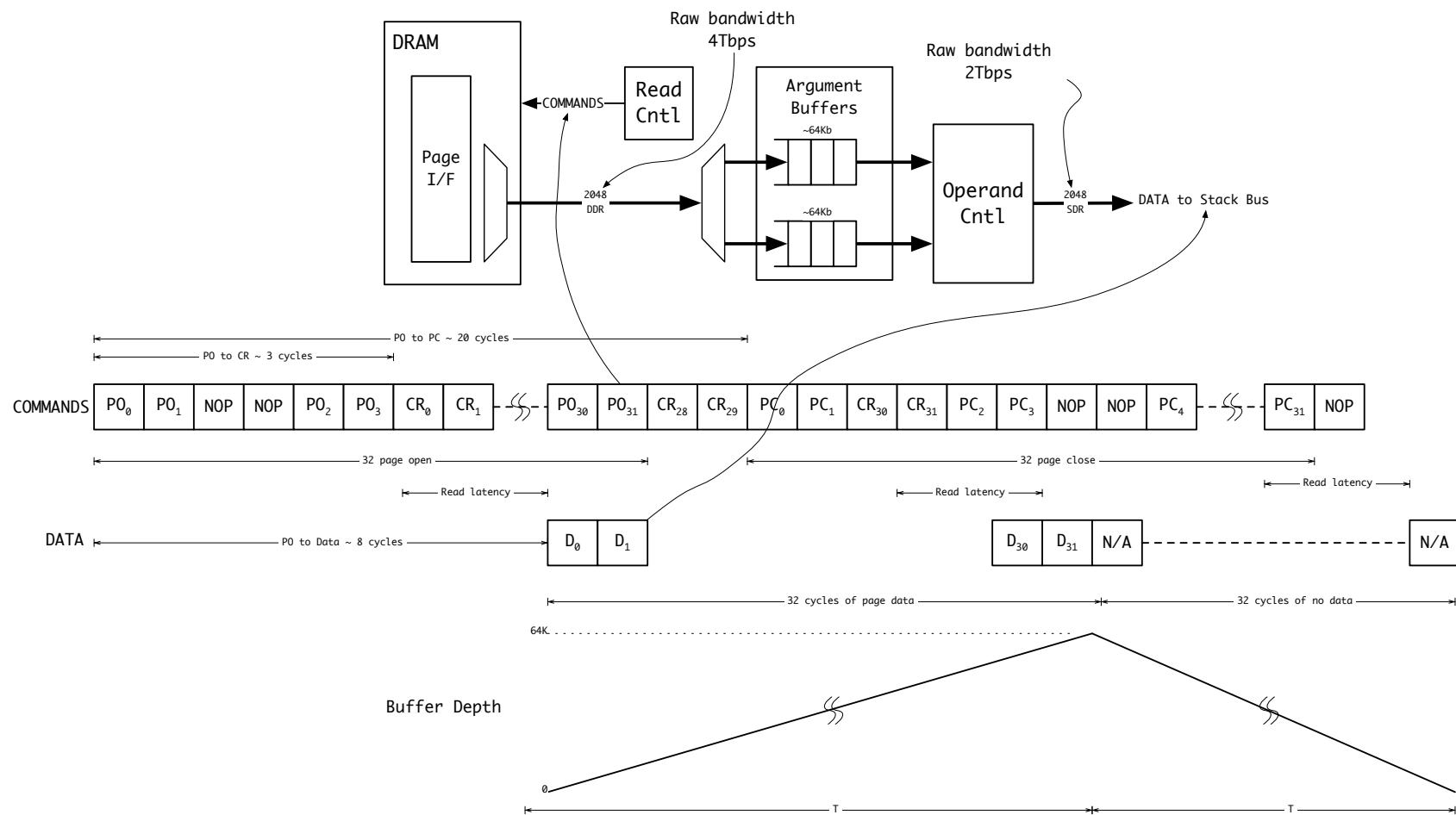
- Additional processing absorbed by layer 1 ANe MAC operation

Pooling Implementation

- Pooling performed by ordering previous layers calculation



Example DRAM Access Sequence



Example Image Recognition CNN

- Example taken from Krizhevsky, Sutskever, Hinton 2012 [Kri12]
- Seven layers but only first 3 layers shown
- ~60M parameters(~2Gb), ~2GFLOP

TSV Power

from [Liu12]

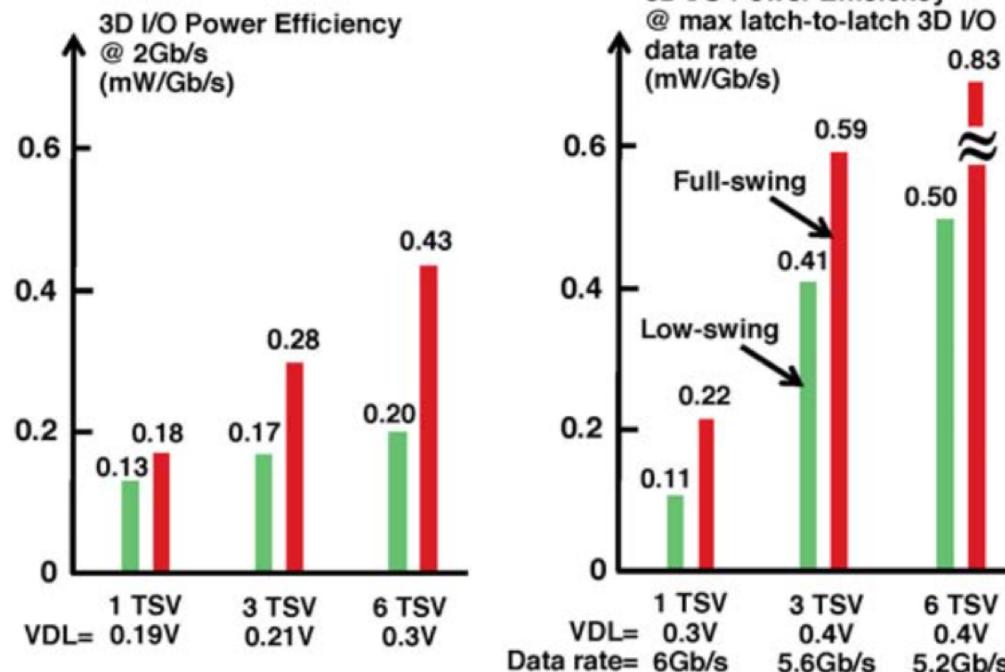


Figure 7.7.5: Measured power efficiency comparison of low-swing and full-swing latch-to-latch 3D I/O test sites.

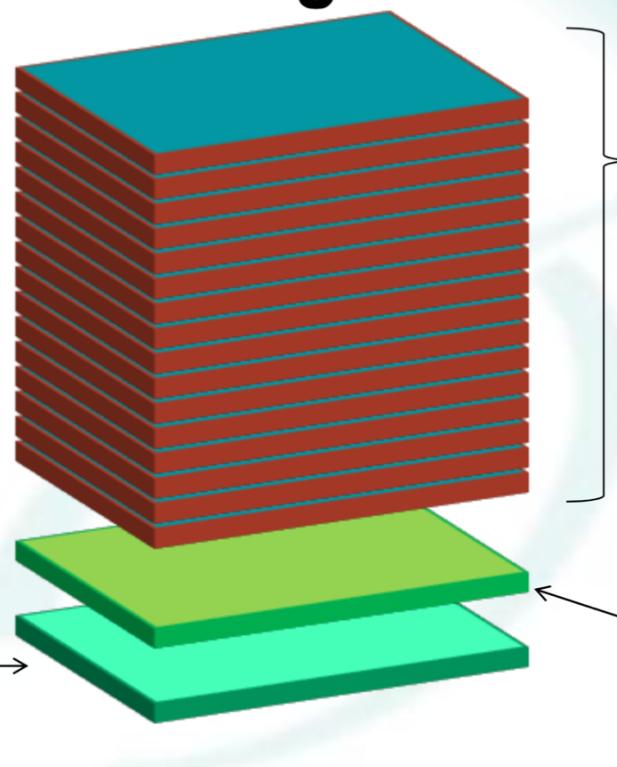
from [Pat14]

DiRAM4 Stack

DiRAM4 “Dis-Integrated” 3D Memory

2 million
vertical
connections
per layer per
die

I/O layer contains:
I/O, interface logic
and R&R control
CPU. 65nm node



Bit Cell layers
4xnm node

Controller layer
contains: senseamps,
CAMs, row/column
decodes and test
engines. 40nm node

FMA Power/Area Estimate

- Consider an FMA running at the Stack bus speed of ~1GHz, from [GH11] table 1

Fused Multiply-Accumulate Power and Area [4]

Pipeline	Freq (GHz)	Area (μm^2)	Static Power (mW)	Dynamic Power (mW)	$W/GFlop$	$mm^2/GFlop$	Power Density (W/mm^2)
6	2.08	16077	1.2	30.9	0.0077	0.0039	2.00
6	2.08	16077	1.2	30.9	0.0077	0.0039	2.00
5	1.32	14241	0.55	12.85	0.0051	0.0054	0.94
4	0.98	12670	0.58	8.09	0.0044	0.0065	0.68
3	0.5	12117	0.16	3.16	0.0033	0.0121	0.27
3	0.2	10619	0.0358	0.952	0.0025	0.0265	0.09

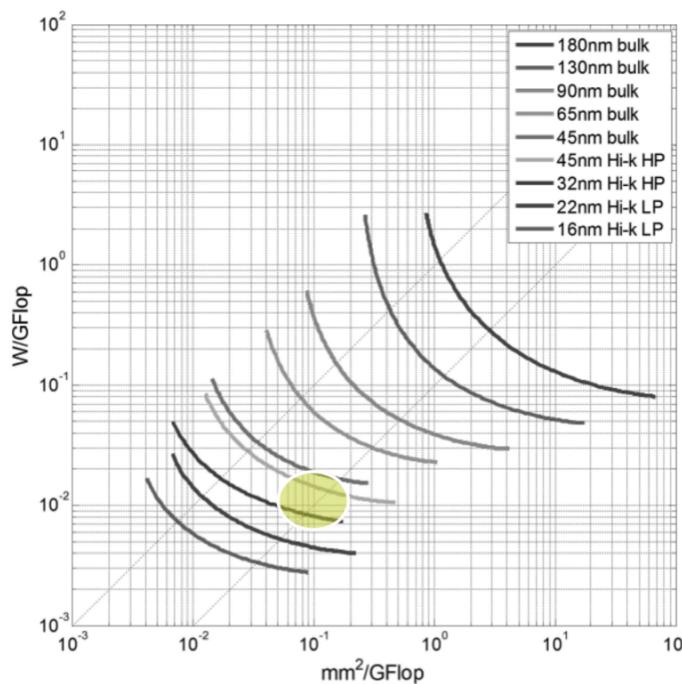
$$\begin{aligned} \text{FMA Area (45nm)} &= \text{area}/fma \cdot \#PE \cdot fma/PE \\ &= \frac{12670}{10^6} \cdot 64 \cdot 32 = 25.95 \text{ mm}^2 \end{aligned}$$

$$\begin{aligned} \text{Max FMA Power (45nm)} &= \text{Number of FMA} \cdot (\text{leakage power} + \text{dynamic Power}) \\ &= 64 \cdot 32 \cdot \left(0.58 + \frac{8.09}{0.98} \cdot 1.0 \right) = 1.187 + 16.91 \\ &= 18.1 \text{ W} \end{aligned}$$

$$\text{Actual FMA Power (45nm)} \approx \frac{54\text{Tbps}}{131\text{Tbps}} \cdot 16.91 + 1.187 = 6.97 + 1.187 = 8.16 \text{ W}$$

FMA Power/Area Estimate

- From [GH11] figure 10 scaling from 45nm to 32nm whilst maintaining the mm²/Gflop suggests a scaling factor of 0.53



$$\text{Actual FMA Power (32nm)} \approx \frac{0.8}{1.4} \cdot 8.16 = 4.66 \text{ W}$$

$$\text{FMA Area (32nm)} = 25.95 \text{ mm}^2$$

TSV Area

- Power Delivery
 - ~2000 TSV's ~0.1mm²
- Stack Bus
 - 2737 signals
 - Assume 2:1 for GND/VCC ~ 5500 TSV's
 - 5um pitch ~ 0.14mm²/PE
 - 8.8mm² for 64 PE's
- DRAM Bus
 - 4255 signals ~ 8500 TSV's
 - 13.6mm² for 64 Ports

State of the Art Older ASIC's

- NeuroCube¹
 - uses HMC 3D-DRAM but dependence on SRAM limits support to CNNs
- Eyeriss²
 - discusses DRAM accesses but their focus is on the convolution operation
 - supports CNN convolution operation only
- NnSP³
 - uses cache between DRAM and 8Mb SRAM and will be limited by DRAM bandwidth
 - does not discuss multi-devices
- Neuflow⁴
 - limits their support to CNNs and in the case of Eyeriss only supports the convolution
 - will not support locally-connected ANNs

¹ [Kim16], ² [Che16], ³ [Esm05], ⁴ [Far11]

State of the Art GPU's

- GPU's target the general market and employ infrastructures not necessary for NN implementation
 - we associate power of GPU solution ~100-200W
 - Will still be limited by memory bandwidth although they are starting to utilize 3D-DRAM
 - general purpose architecture makes it hard/impossible to make full use of theoretical performance

	CPU	V6	mGPU	IBM	GPU
Peak GOPs	10	160	182	1280	1350
Real GOPs	1.1	147	54	1164	294
Power W	30	10	30	5	220
GOPs/W	0.04	14.7	1.8	230	1.34

Table 5. Performance comparison. 1- CPU: Intel DuoCore, 2.7GHz, optimized C code, 2- V6: neuFlow on Xilinx Virtex 6 FPGA—on board power and GOPs measurements; 3- IBM: neuFlow on IBM 45nm process: simulated results, the design was fully placed and routed; 4- mGPU/GPU: two GPU implementations, a low power GT335m and a high-end GTX480. [Far11]

Detailed Block Diagram

