Mythic's Analog Deep Learning Accelerator Chip: High Performance Inference

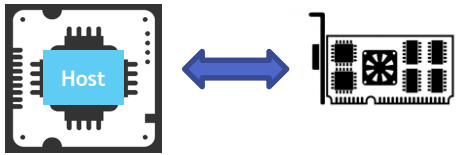
MYTHIC

Frederick Soo, Ph.D.

May 22, 2018

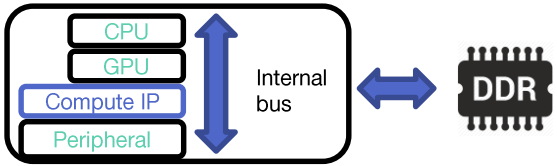
Poor options for high performance embedded deep learning

External GPU



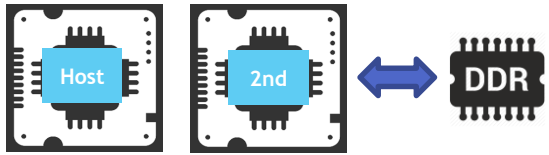
- High power (100W+) weight (1kg+)
- High CAPEX and OPEX

IP core



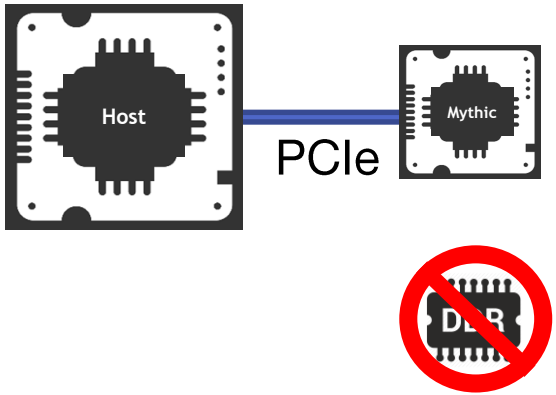
- ASIC cost and risk
- Pressure on DDR memory buses
- Current solutions not powerful enough for many applications

Secondary processor



- Complexity of second processor
- Does not avoid DDR and power requirements

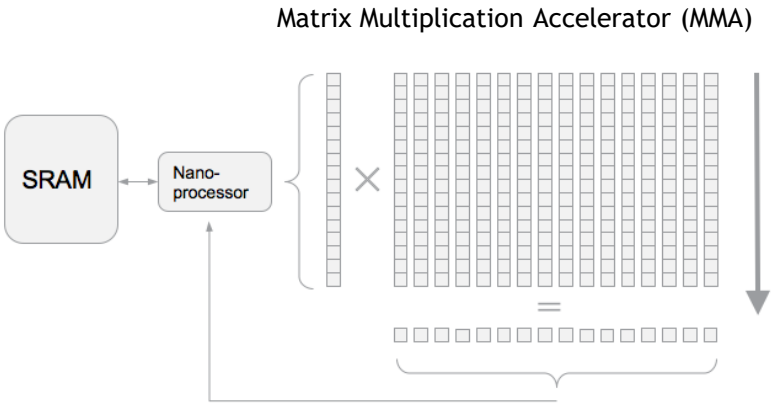
Mythic solution – single chip analog DNN



- **DNN processor** equivalent to 1 nVidia Titan XP but at less than 2W power
- **Single chip** can handle Full HD workloads at low latency
- **Multiple chips** can be hosted on PCIe bus
- **Deterministic execution** – statically compiled program, fixed scheduler, guaranteed timing and power
- **No DDR** needed to support chip
- **Low CPU overhead** – data transfers through PCIe DMA

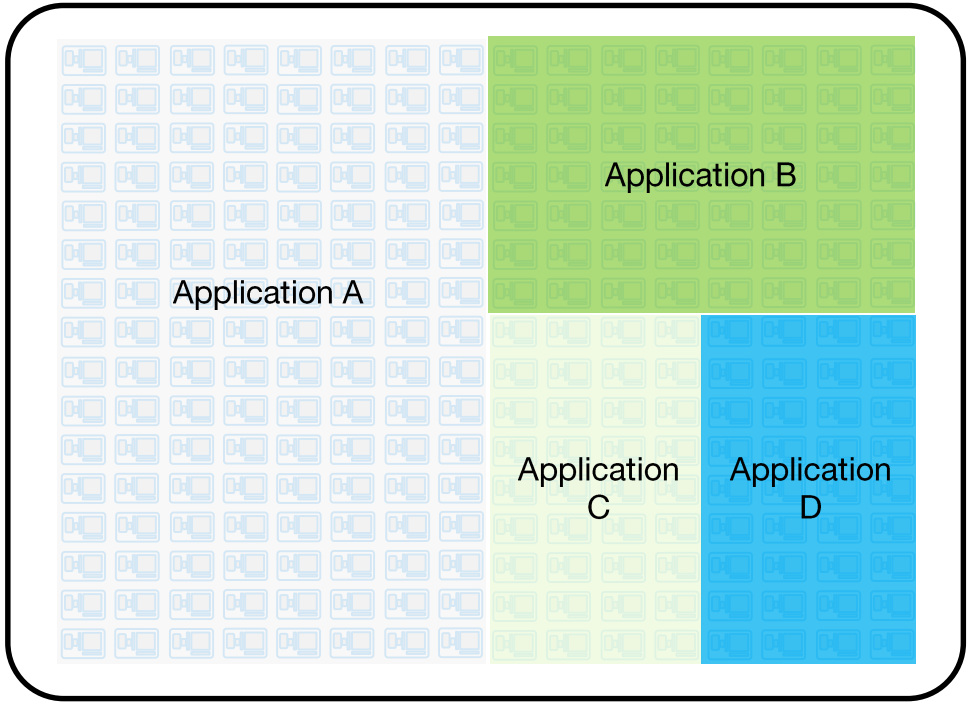
Mythic analog IP

Mythic compute tile



- Research began in 2012 with first DNNs
- Weights stored on flash transistors and multiply done in analog domain – 50x denser than SRAM.
- Best approach out of RRAM, MRAM, weight compression, etc. – noise, stability, compatibility, process
- Partnership and support from major flash memory vendor
- Analog block validation in 2018, with final verification in mid 2019.
- Room for scaling - at 40nm can beat current 10nm digital with room for 4-5x process improvement

Mythic chip floorplan



- 50-100 tiles connected by programmable digital network
- Independent compute for multiple applications on a single chip
- Deterministic execution for safety and timing critical applications
- Power gating of tiles to increase efficiency

Example: ResNet-50 (atrous)

Mythic estimated performance vs DSP and GPU

Configuration	GMAC/sec	Resolution	Inferences/sec	Average # of active tiles[a]	Power[c] (W)	# of weights (including parallel copies)	pJ/MAC	TOPS/W
Mythic	116	224x224	30	3	<0.5	25M	0.6	3.3
	3480	224x224	900	98	<2	33M		
	700	640x480[d]	30	20	<0.5	25M		
	4800	1920x1080[d]	30	135[b]	<3	41M		
SoC DSP[e]	29	224x224	7.6	-	0.5	N/A	15	0.13
SoC DSP[e]	80	1920x1080[d]	0.5	-	0.7	N/A	8.8	0.13
Titan XP[f]	3450	224x224	892	-	167	N/A	48	0.04

[a]Generation 1 chip: 50 tiles @ 1M weights/tile.
[b]Network would be split across three Generation 1 Mythic chips
[c]Mythic estimates; actual dynamic measurements of DSP and GPU using power meter. Does not include CPU+peripherals+static power of DSP (about 1W for DSP) [f]TensorRT INT8; could not generate Full HD INT8 engine (out of memory)

[d]Fully convolutional version of ResNet-50 – *fc1000* 1x1 is applied to last convolutional layer, followed by global average pool
[e]INT 8 runtimes generated using SDK provided by manufacturer
[f]TensorRT INT8; could not generate Full HD INT8 engine (out of memory)

~15-30x more efficient than SoC-based INT8 DSP

Comparable to Titan XP but at 2W power

4 copies of ResNet-50 @ 640x480 30fps can be run on a single chip simultaneously

PCIe power consumption is low

Resolution	Raw data rate[a] (gbps)	PCIe 2.1 lanes[b]	% usage	Expected PCIe PHY power[c]
VGA	0.2	1	5.5%	**<10mW**
Full HD	1.5	1	38%	**50-70mW**
4K	7.0	2	87%	**200mW**

[a]30fps, 24 bit RGB
[b]PCIe Gen 2.1: 4gbps/lane
[c]100mW/lane; VGA - L2 low power mode > 90%; HD - mix of L1 and L2 modes

- Less than 10mW for light loads
- Small power footprint compared to sensor (100mW), SoC +memory (0.5-1W+) and other parts of system (LTE, WiFi)
- 4K supported with 2 PCIe Gen 2.1 lanes

Example: SSD with ResNet-18 front end[a]

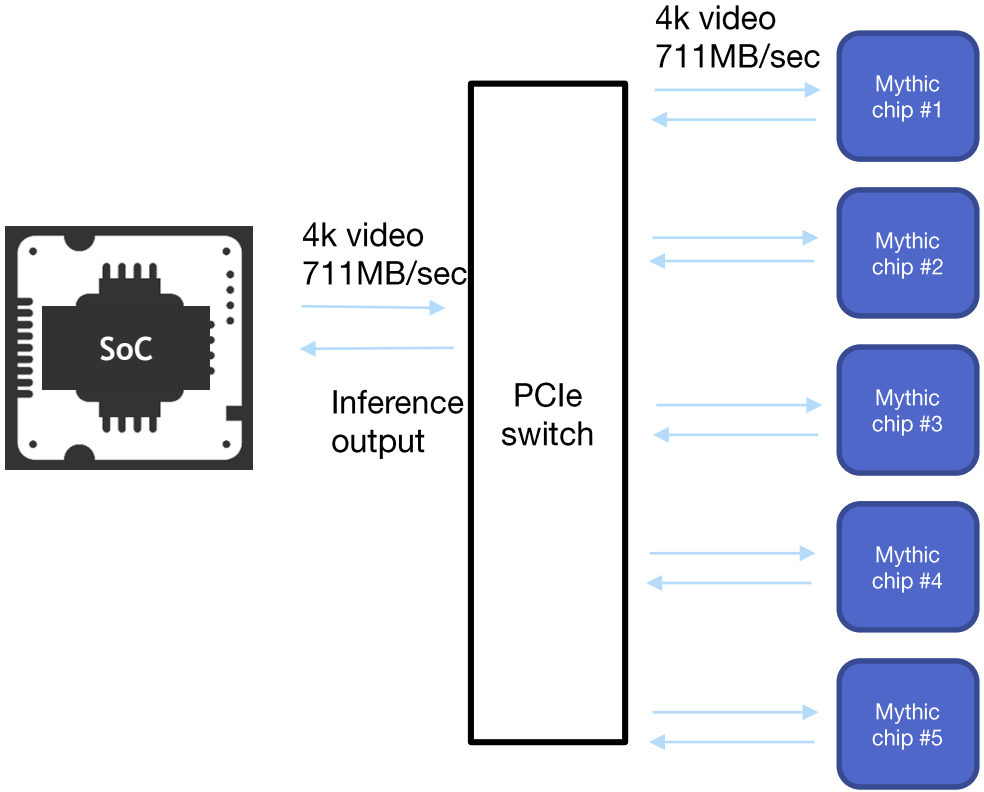
Mythic estimated performance

Resolution	Inferences /sec	Input bandwidth (MB/sec)	Average # of active tiles[b]	Power[c] (W)	# of weights (including parallel copies)	# of chips
300x300	60	15	5	<**0.5**	16M	1
VGA	60	53	20	<**0.5**	16M	1
Full HD	30	178	60	<**1.5**	26M	2
4K	30	712	225	<**5.0**	64M	5

- Can scale to Full HD @ 30fps, 1.3W, single chip
- Multi-camera, 4K can be handled by multiple chips at <5W
- Millisecond-range latency
- Does *not* include non-max suppression

[a]SSD with atrous architecture – VGG-16 FC layer causes blow-up at higher resolution
[b]Mythic chip Generation 1 - 50 tiles with 1M weights/tile
[c]Estimated power

Example: SSD[a] 4K multi-chip @ 30fps



- Chip-to-chip daisy-chain
- Each chip handles a subset of neural network layers
- Each chip has 4 lanes, can handle 2GB/sec bidirectional transfers
- 4K video @ 30fps by 5 chips
- More chips to increase throughput, reduce latency, run larger networks
- Low latency – some additional overhead from PCIe but still in low milliseconds

Example: OpenPose with ResNet-18 front

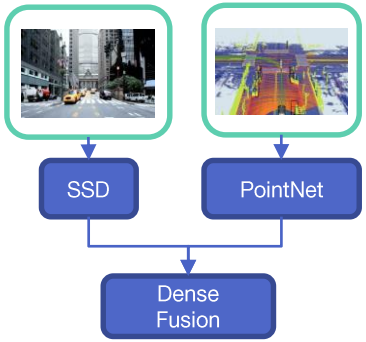
Mythic estimated performance vs DSP and GPU

	Resolution	Inferences/sec	Average # of active tiles[a]	Power[b] (W)	# of chips[c]	pJ/MAC	TOPS/W
Mythic	328x176	60	26	<1	1	0.2	10
	656x368	60	105	<2.5	2		
	1920x1080	30	444	<10	5		
SoC DSP[d]	656x368	0.7	-	1	-	8.1	0.24
Titan XP[d]	656x368	68.5	-	225	-	18.3	0.1

[a]Convolutional network only, ResNet-18 front end – does not include video pipeline and post-processing
[b]Estimated power; DSP and TitanXP power measured using power meter and nVIDIA-SMI, respectively
[c]Generation 1 chip has 100 tiles planned. Use multiple chips to increase throughput and lower latency
[d]INT 8 runtimes

- 50-100x more efficient than DSP or GPU running TensorRT-Int8
- Can run QVGA 60fps @ <1W
- Can scale to FHD @ 30fps @ <10W

Example: Video+LIDAR fusion



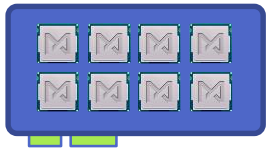
Mythic estimated performance

Component	Resolution	Rate (fps)	Average # of active tiles	Power[a] (W)
PointNet[b]	300 points x 400 objects	10	54	<1.5
Dense Fusion	300 points x 400 objects	10	22	<0.5
SSD[c]	Full HD	10	19	<0.5
Total			**95**	**<3.0**

- Full HD @ 10fps video+LIDAR fusion onto single chip at 2W
- Scale system for higher resolution, frame rate and multiple cameras by adding chips
- Does not include non-max suppression for SSD

[a]Estimated power
[b]Spatial transform network replaced by camera rotation as in PointFusion paper (https://arxiv.org/pdf/1711.10871.pdf)
[c]SSD with ResNet-18 front end

Example: Multi-chip PCIe card



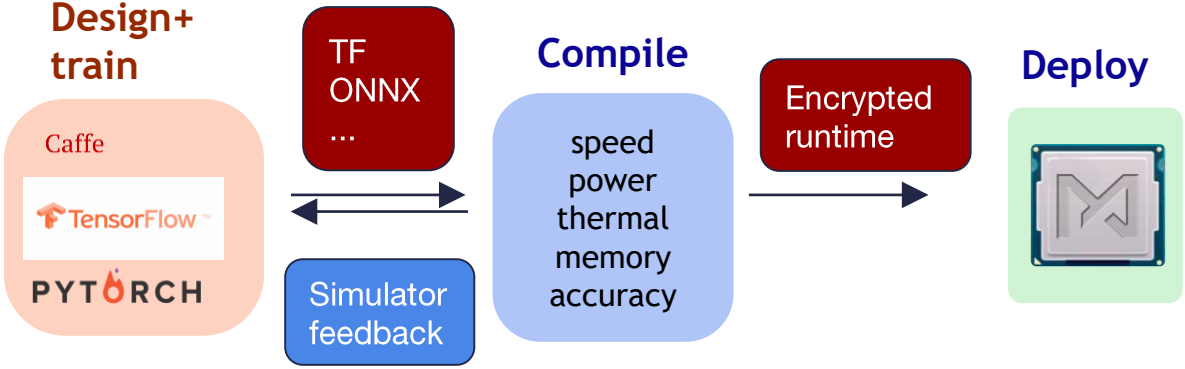
Example: ResNet-50[a] @ 1920x1080 @ 30fps/chip

# of chips on card	Inferences /sec	Input bandwidth (GB/sec)	PCIe 2.1 lanes	PCIe 4.0 lanes
1	30	0.2	0.5	0.1
8	120	1.0	1.9	0.5
32	480	3.8	7.6	1.9

[a]Fully convolutional version of ResNet-50 – *fc1000* 1x1 is applied to last convolutional layer, followed by global average pool

- Deep networks are not I/O limited
- Most of processing power used on deep layers
- Small fraction of total tile compute is dedicated to input
- For ResNet-50 and similar networks, 4 lanes of PCIe Gen 2.1 sufficient to feed ~16 Mythic chips
- 16 lanes of PCIe 4.0 can feed ~128 Mythic chips

Production-optimized workflow



Tensorflow current target, will also support ONNX and other frameworks

Statically compiled programs – deterministic runtime

Encrypted runtime and secure API

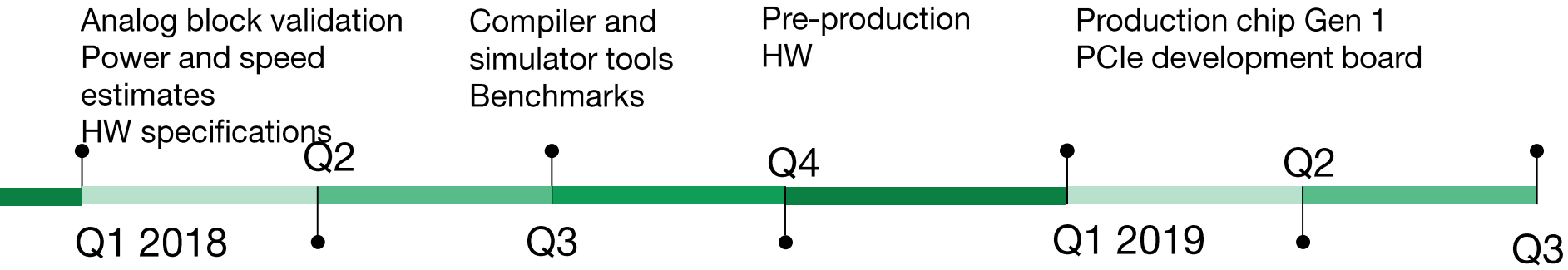
Simple driver API

How Mythic will get to market

- Highest risks – analog design – have been mitigated/proven
  - Third generation of analog validation chip
  - Analog performance - programming, read, write, power, noise – within spec
  - *Very low analog power baseline allows us to trade power/noise, offload to digital if necessary*

- Hard but solvable parts are ahead – digital, software, driver, algorithm:
  - Digital design – inter-tile networks, nano-processor, I/O, debug
  - Compiler – scheduling, optimization
  - Drivers and software – PCIe, host-side APIs
  - Algorithms – quantization and noise
  - *Problems here can be solved through software and firmware*

Full production in 12 months



Early adopters

- We are looking for a small set of early adopters:
  - Sophisticated about neural networks
  - Stringent latency, power, throughput and form-factor requirements
  - High value in performance –increased accuracy, speed/power/latency
  - Can prototype/test/deploy quickly (e.g. using Jetson TX2 or QC already, can use PCIe or miniPCIE cards)

Developer community

- We want to make sure our chips work for developers:
  - Most important models for production
  - Most relevant benchmarks
  - Performance requirements (power, latency, accuracy)
  - Hardware requirements (target SoC, power, form factor)
  - Software requirements (driver, API)
  - Accelerators to market (development kits, reference designs, prototypes)

Take-aways

- 1 Mythic chip ~ Titan XP but at 2W!
- Compatible with Tensorflow and other frameworks
- Scalable by using multiple chips
- Proven analog design
- Launching next spring with PCIe plug-in board
- Work closely with early adopters in 6-12 months on software and integration

Final note

- We are hosting a celebration of our 40M Series B in our Redwood City office!
- Thursday May 31, 2018
- Please RSVP at https://mythic.splashthat.com