Optimized FPGA fabric for Edge Compute and ML Inferencing

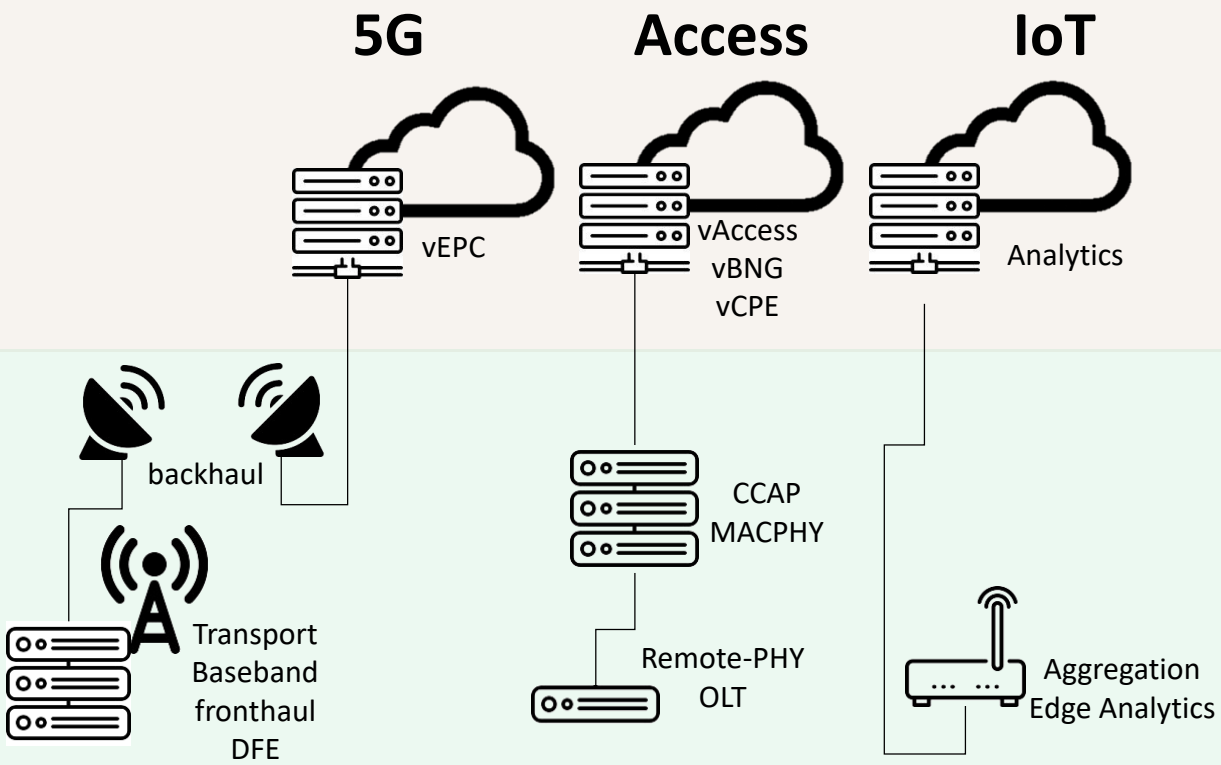
DAC2019

Mike Fitton, Achronix

Achronix®

Data Acceleration

Evolving Requirements for end-to-end Network



Cloud / Data Center

- Significant increase in computation for Inference and Training
- Memory bandwidth as important as computation blocks
- High rate connectivity requirement

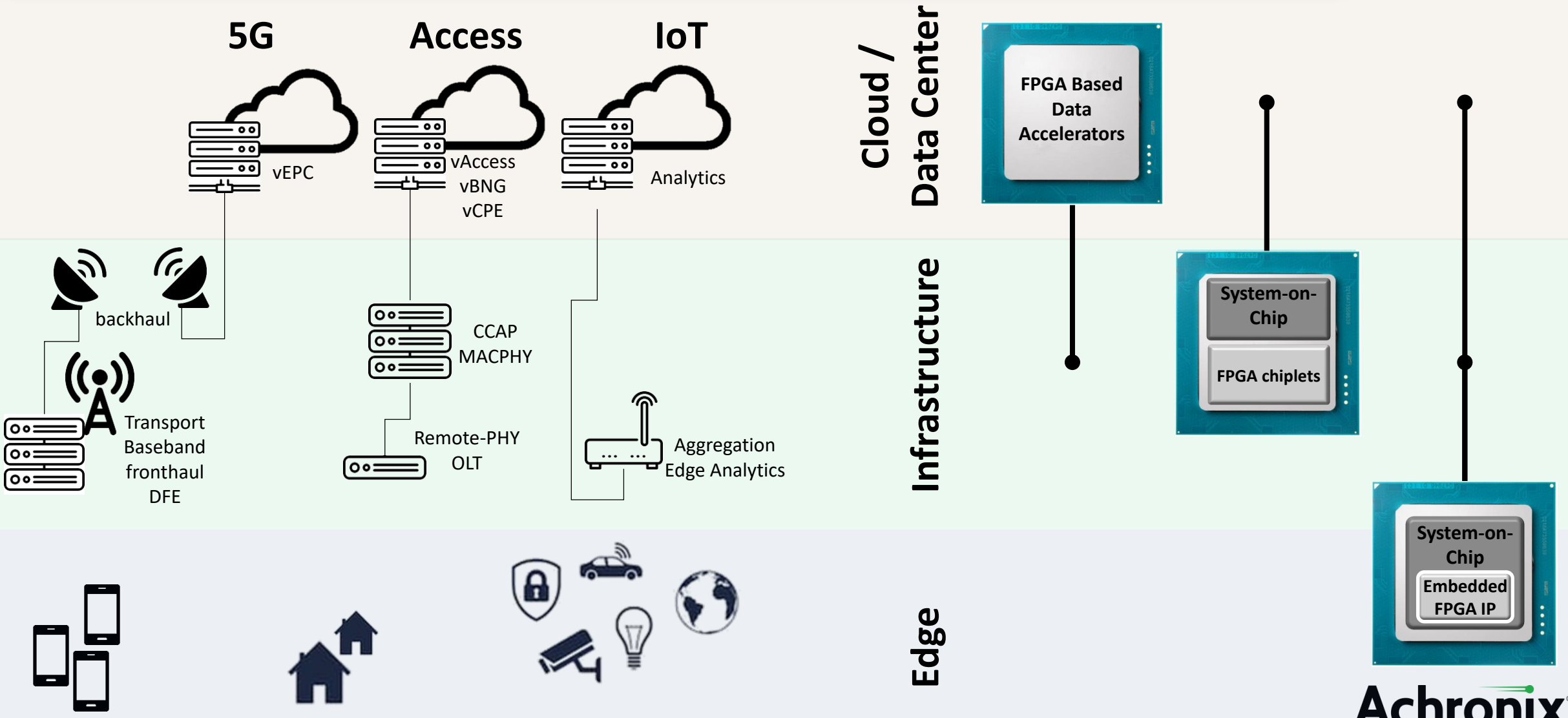
Infrastructure

- Significant increase in computation and moving compute closer to data
- Emergence of Mobile Edge Compute
- Flexibility and futureproofing in algorithms and applications, e.g. in AI/ML arithmetic, 5G, etc

Edge

- Advanced security for all data sets
- Low-cost/power requirement
- Analytics in the end-point

Applicability of FPGA from Data Center to Edge to IoT



Achronix
Data Acceleration

Critical Architecture Requirements for Efficient Data Acceleration

Compute

Data Transfers

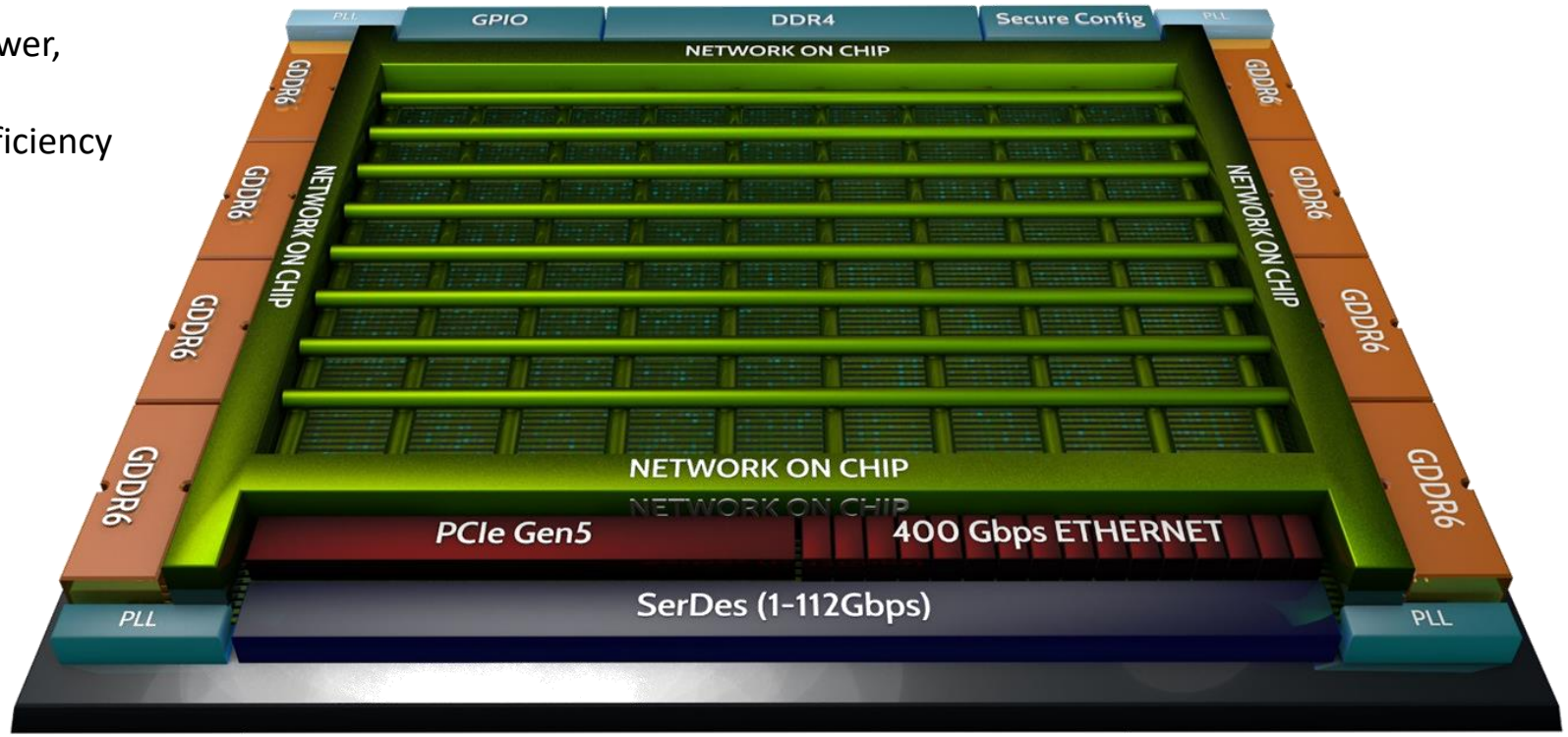
Memory Hierarchy

Objective: Deliver highest performance/watt with adaptability

Achronix
Data Acceleration

Introducing Speedster7t: FPGA Devices with ASIC Performance

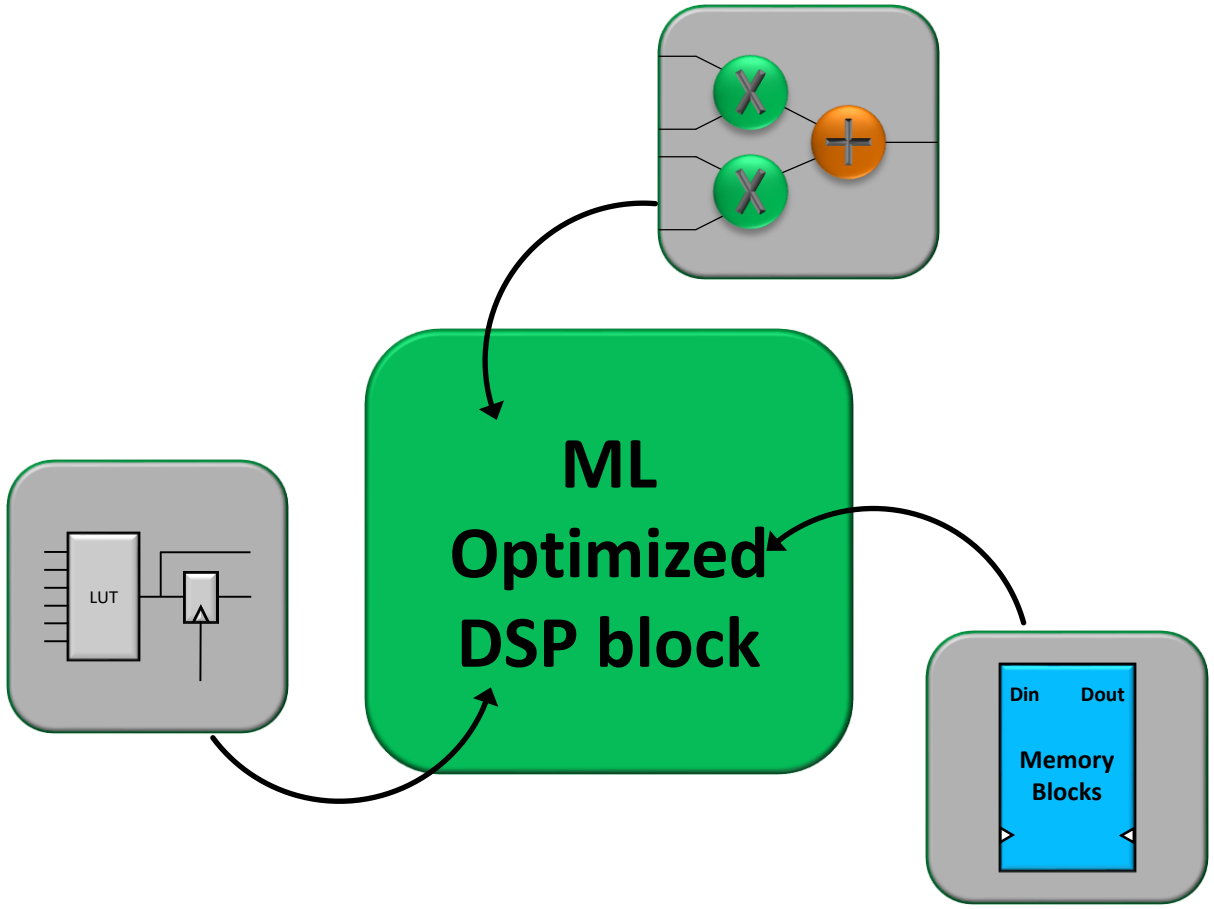
- Speedster7t is a new class of FPGA optimized for high bandwidth workloads
  - Leverage FPGA architecture for optimal power, performance and flexibility
  - Break historical bottlenecks that reduce efficiency
  - ASIC performance in FPGA architecture
- Highest efficiency **Compute** engines
  - New Machine Learning Processor (MLP) gives programmability for e.g. new networks and number formats
- Efficient **Data Transfers**
  - New Network-on-Chip eliminates routing congestion
  - Full flexibility with FPGA byte- and bit-wise routing
- Balanced **Memory Hierarchy**
  - Highest speed interfaces and external memory bandwidth



Achronix
Data Acceleration

Optimizing DSP block for ML

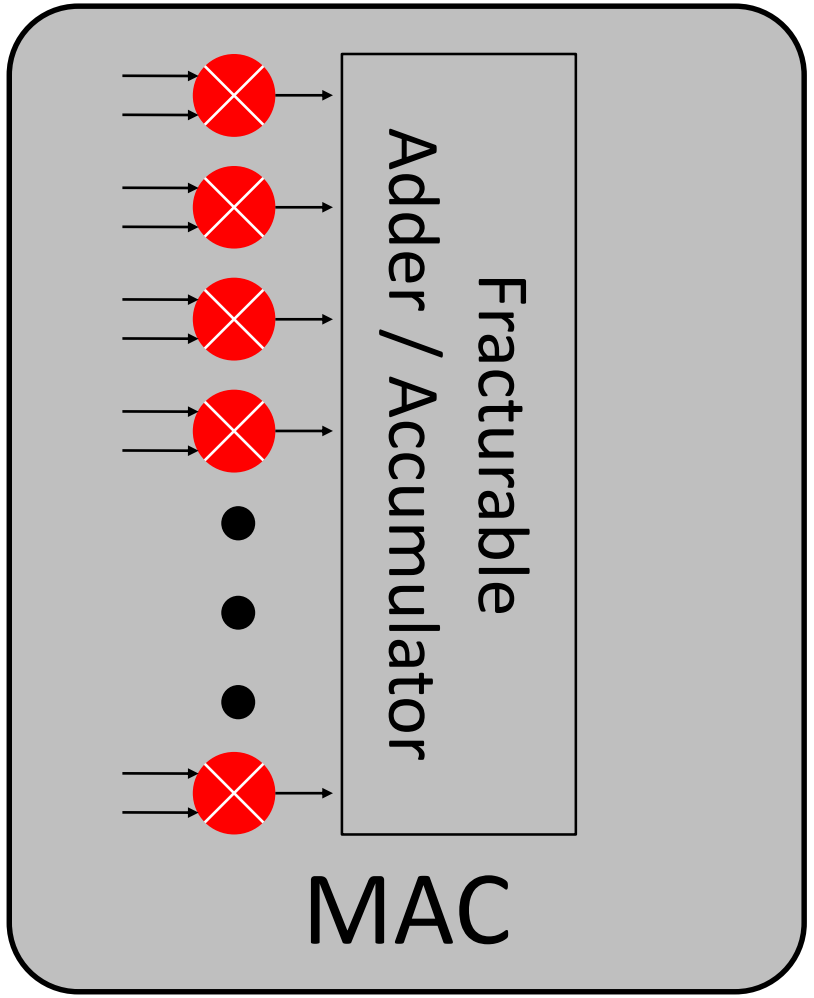
- Traditional FPGA DSP blocks are optimized for filtering
  - Less efficient for lower bit width formats
  - Consumes LUTs/memories to build AI/ML applications
  - Performance limited by FPGA routing
- Machine Learning Processor (MLP) is an ML focused Multiply-Accumulate macro
- Math, memory and programmability all in the MLP block for AI/ML and high bandwidth compute algorithms:
  - Array of MAC compute structures
  - Tightly coupled memories
  - Configurable to support changing algorithms
  - Supports widest range of integer and floating numerical formats



Achronix
Data Acceleration

Math Block Optimized for AI/ML

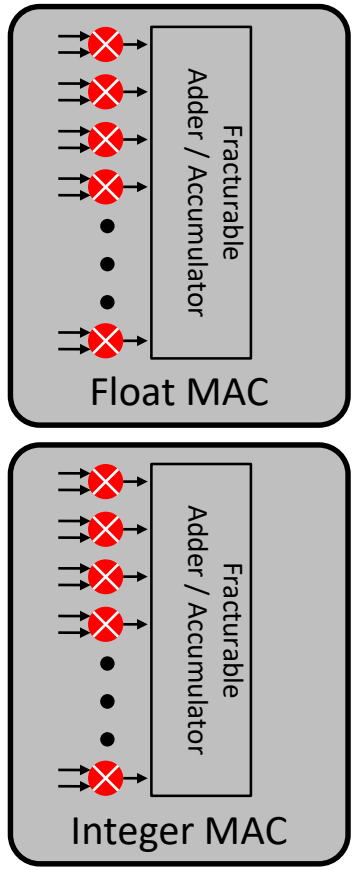
1. High density multiplier arrays
   - Up to 32 multipliers per MAC block
   - Drive variable precision adder / accumulator



Achronix
Data Acceleration

Math Block Optimized for AI/ML

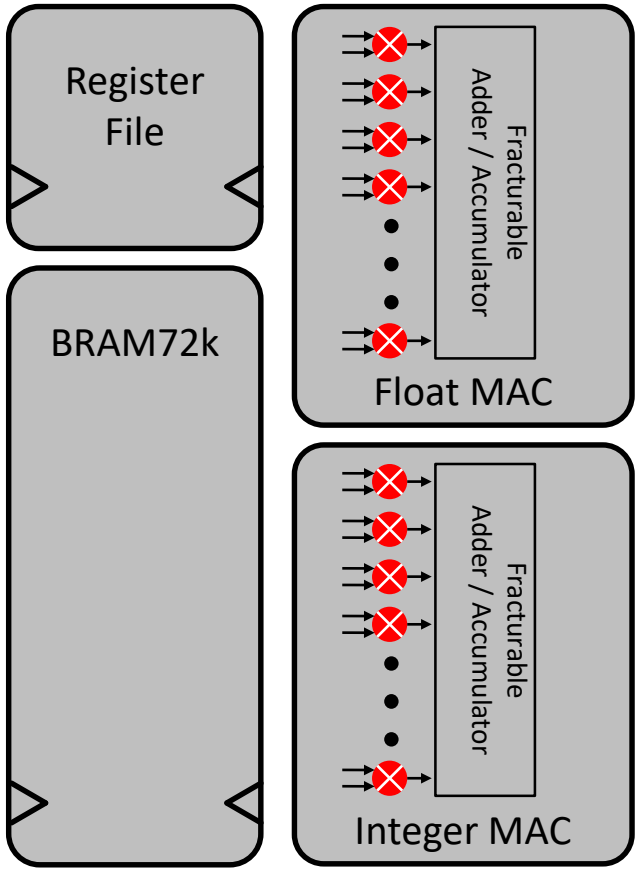
1. High density multiplier arrays
   - Up to 32 multipliers per MAC block
   - Drive variable precision adder / accumulator
   - Floating point MAC and Integer MAC



Achronix
Data Acceleration

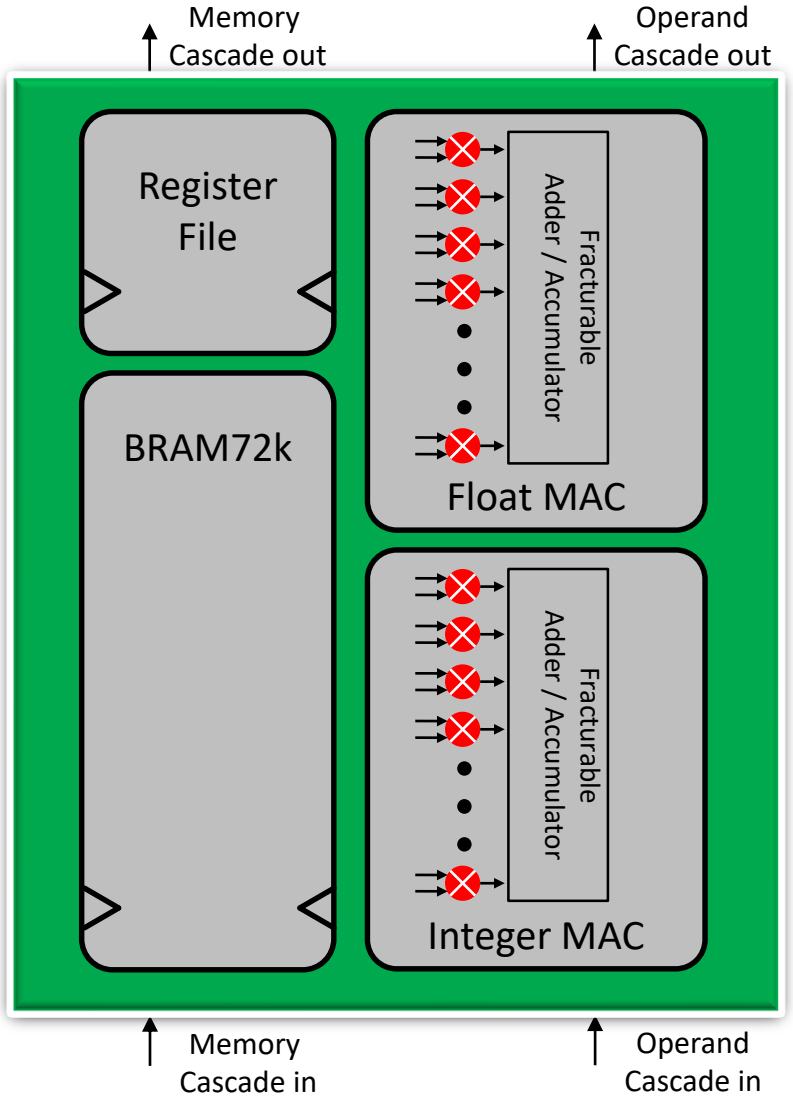
Math Block Optimized for AI/ML

1. High density multiplier arrays
   - Up to 32 multipliers per MAC block
   - Drive variable precision adder / accumulator
   - Floating point MAC and Integer MAC
2. Tightly coupled memory blocks
   - Large block RAM (e.g. 72Kbits)
   - Register File (e.g. 2K bits)



Math Block Optimized for AI/ML

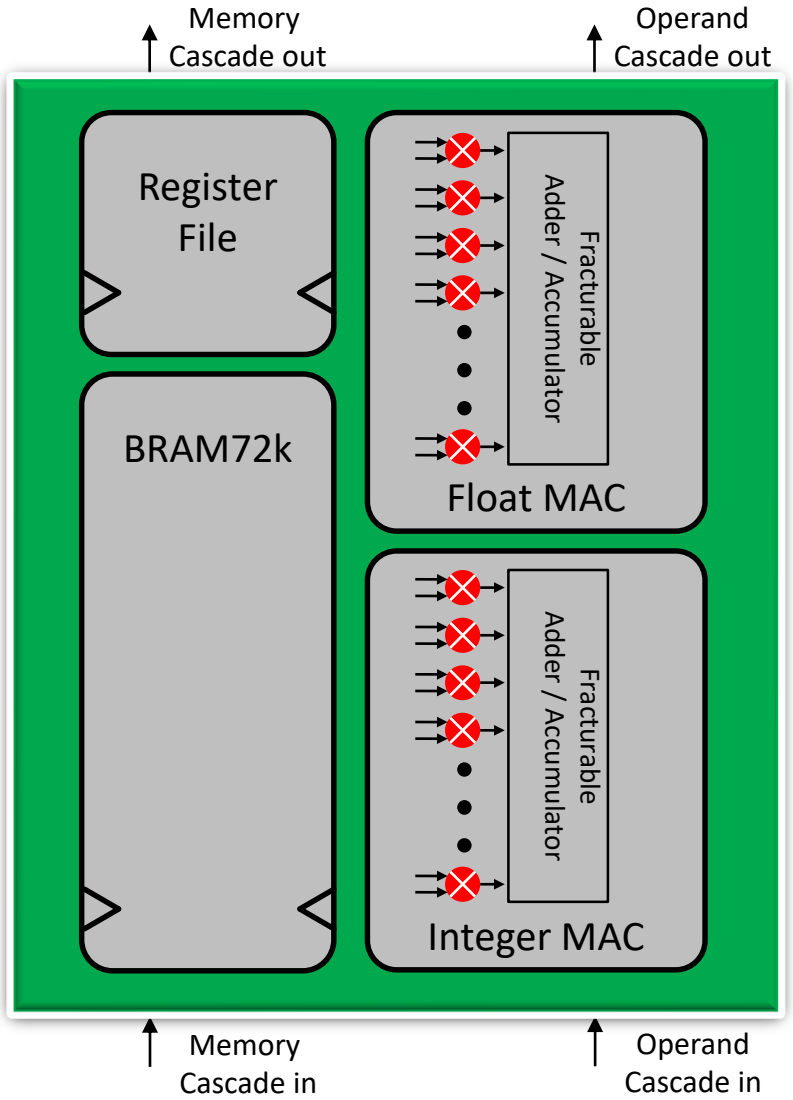
1. High density multiplier arrays
   - Up to 32 multipliers per MAC block
   - Drive variable precision adder / accumulator
   - Floating point MAC and Integer MAC
2. Tightly coupled memory blocks
   - Large block RAM (e.g. 72Kbits)
   - Register File (e.g. 2K bits)
3. Operand and memory cascade functions
   - Create larger algorithms without using FPGA routing or resources



Achronix
Data Acceleration

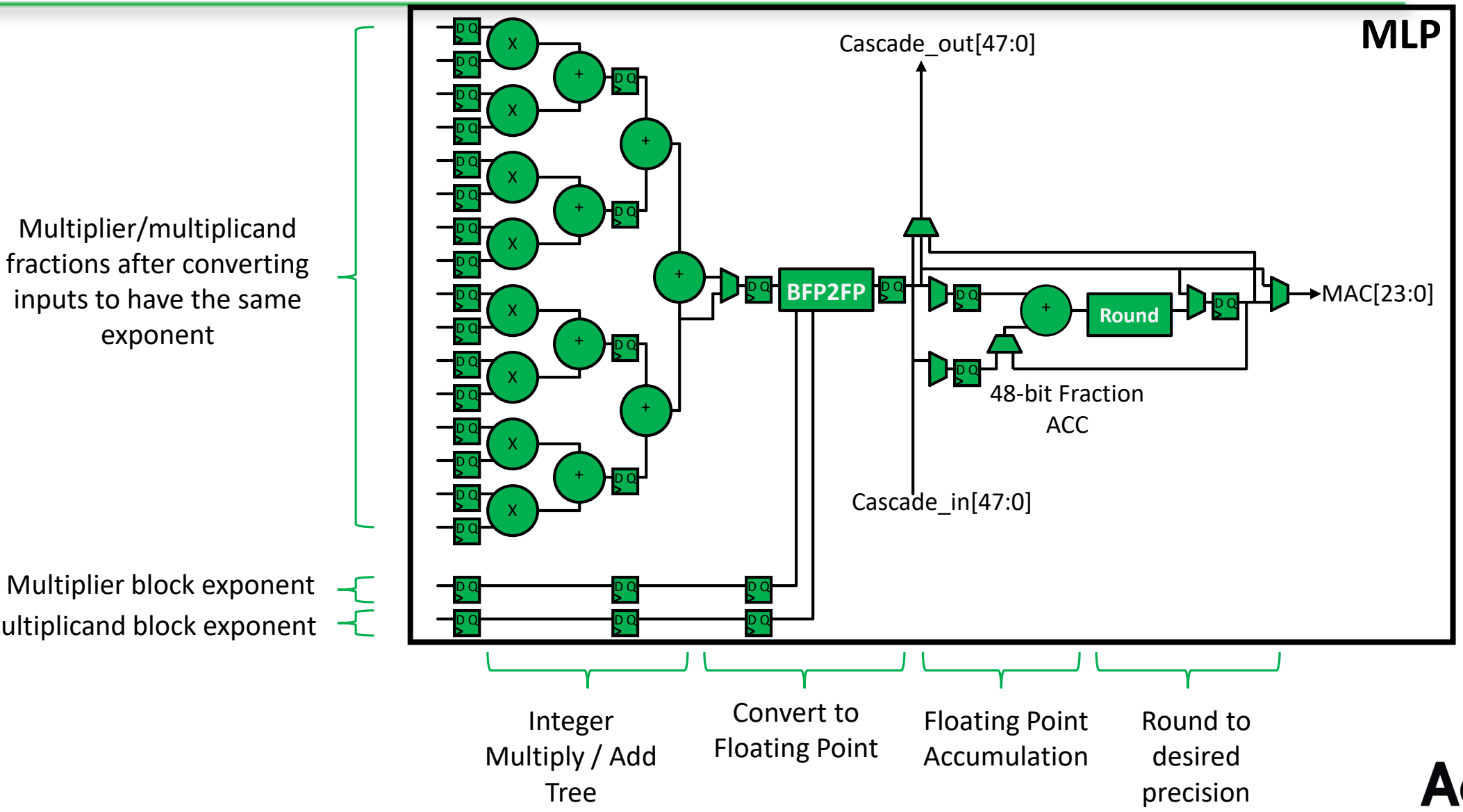
Math Block Optimized for AI/ML

1. High density multiplier arrays
   - Up to 32 multipliers per MAC block
   - Drive variable precision adder / accumulator
   - Floating point MAC and Integer MAC
2. Tightly coupled memory blocks
   - Large block RAM (e.g. 72Kbits)
   - Register File (e.g. 2K bits)
3. Operand and memory cascade functions
   - Create larger algorithms without using FPGA routing or resources
4. Support for multiple number formats
   - Floating point:
     - FP16, bfloat16, FP24
     - Block floating point with shared exponent
   - Integer:
     - Int16, Int8, Int4



Achronix
Data Acceleration

MLP Block Floating Point Mode

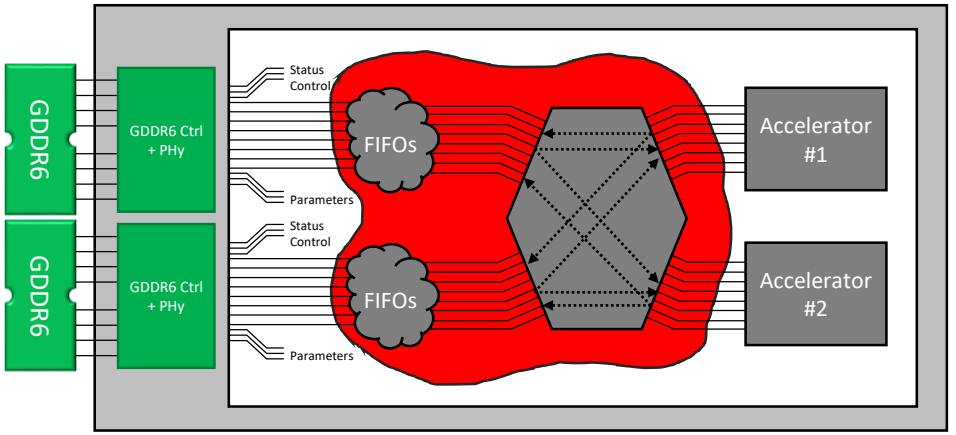


Achronix Data Acceleration

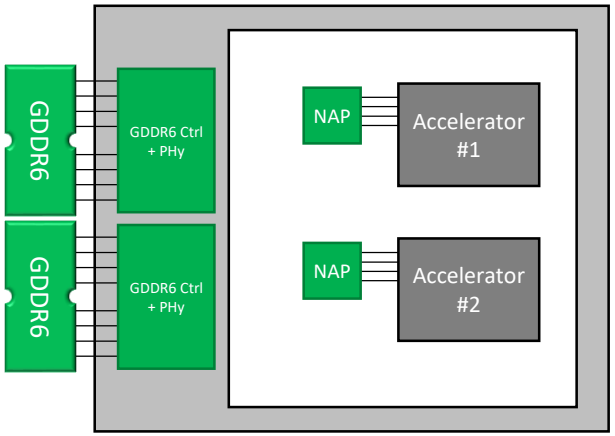
High Bandwidth Network on Chip (NoC) in next generation FPGAs

- High Performance connectivity between Network Access Points (NAPs), interfaces and external memory
  - Each row/column: 256b @ 2.0 Gbps bidirectional
  - Maximum device bandwidth = 20 Tbps
- Example usage models
  - Store PCIe traffic in external memory
  - Preload block RAMs with weights
  - Connect accelerators (NAP to NAP)
- NOC enable high-performance, efficient Data Acceleration
  - Ease-of-use: simple AXI or streaming interface
  - FPGA resources not used for NOC routing
  - Eliminates floorplanning challenges: step-and-repeat

Traditional FPGA Design



Speedster7t FPGA+ Design



Achronix Data Acceleration

Benchmarking examples: Conv2d, Yolo-v2 and ResNet-50

Dimensions batch,in_height,in_width,in_channels filter_height, filter_width, out_channels	Numerical format	Strides	Performance		Resource Usage		
			Throughput	Freq (MHz)	#MLPs	#BRAMs	#LUTs
Conv2d 1,227,227,3,11,11,1	int8	1	1 image in 137us	750MHz	1	5	585
Conv2d 32,227,227,3,11,11,1	int8	4	32 images in 137us	749MHz	32	36	3413

Network	AC7t1500 (images/sec)
ResNet-50	8,600
Yolo-v2	1,600

- 80% MLPs used, 16 mult 8x8 per MLP
- MLP Fmax: 750 MHz

Achronix Data Acceleration

FPGAs for Edge Compute and ML Inferencing

- FPGAs can provide a flexible, reprogrammable workload accelerator for Edge Compute
  - Data interfacing: IO adaption and sensor fusion
  - Pre-processing: format conversion and compression
  - Matrix Vector Maths where flexibility is required in ML
- Performance is maximized with targeted design approach
  - Data Compute
    - Optimize Multiply-Accumulate hard macro for matrix multiplication
    - Support of low bit width integer, block floating point and half-precision float
  - Data Transfers
    - High performance interfaces and dedicated Network-on-Chip
  - Data Memory Hierarchy
    - High bandwidth external memory
    - Distributed memory hierarchy on FPGA
- Achronix Speedster7t is an example of an FPGA optimized for Networking and ML