Introduction to C7x DSP

Automotive Processor Business,

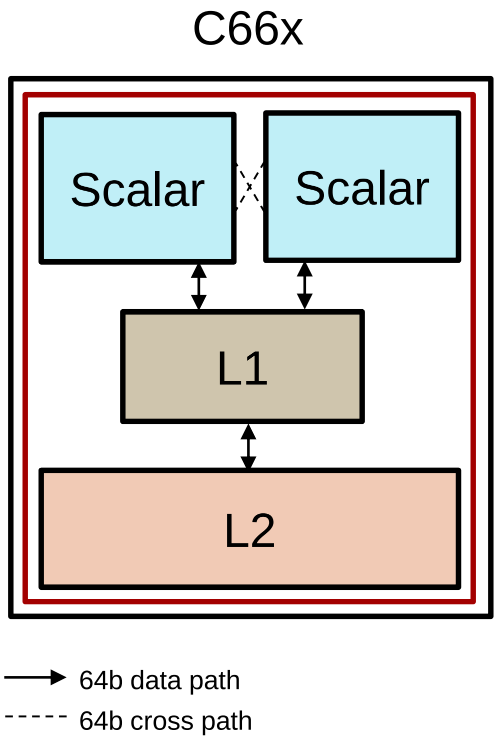
Texas Instruments

Outline

- C66x DSP today
- A bird’s eye view of C7x DSP
- Under the hood of C7x DSP
  - Data path
  - Functional Units
  - Register File
  - Streaming Engine
  - Memory
  - LUT/Histogram
  - ISA

C66x DSP today

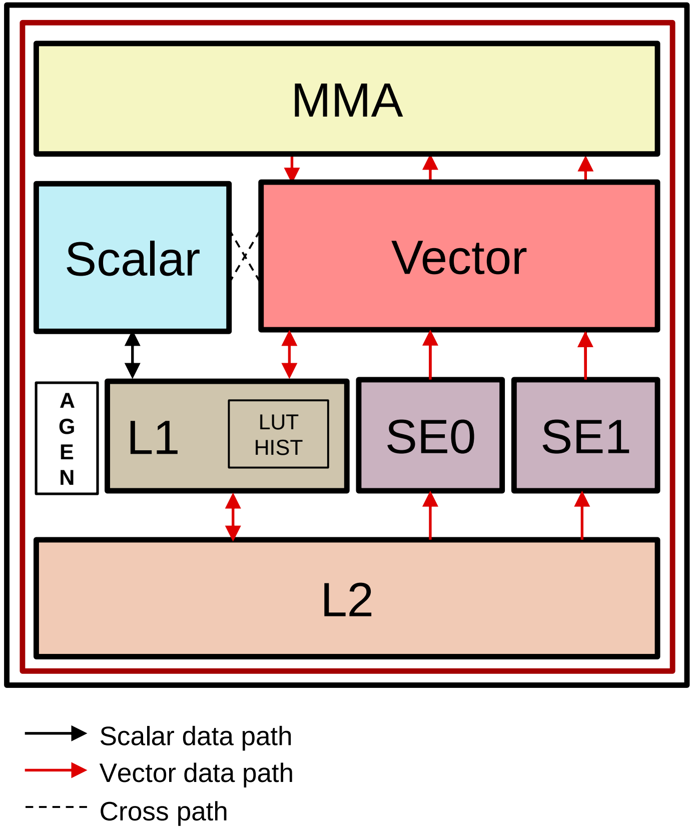
- True 32b floating point DSP
- Programmable functional units (.L1/.L2, .S1/.S2, .M1/.M2, .D1/.D2)
- Global register files (32x2, 32bit registers)
- Cache based memory system (L1D – 32KB, L1P-32KB, L2 – 256KB)
- Dual 64bit data-paths
- Packed SIMD operations (8bit, 16bit, 32bit, 64bit)
- Supports 16bit/32bit complex types
- Supports 128bit vector types (Quad 32bit)
- Supports 32 16bit multiply-accumulate per cycle
- Supports 16 single precision operations per cycle
- Supports 40bit operations
- Software pipelining, special SPLOOP HW



A bird's eye view

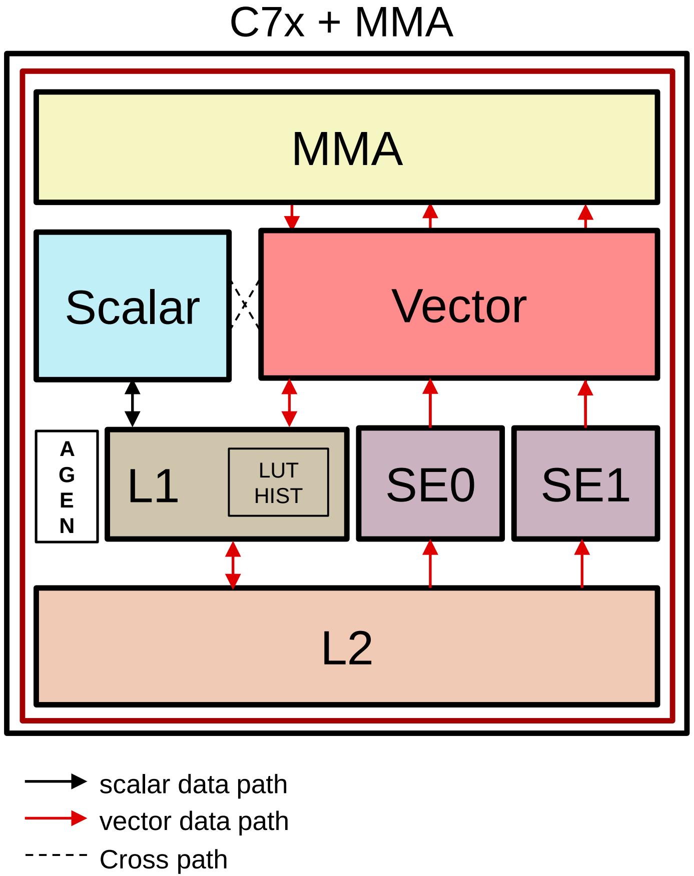
- True 64b DSP with dual data paths
- Programmable functional units
- Global and local register files
- Cache based memory system (L1, L2)
- Streaming Engine (SE0, SE1)
- Address generators (AGEN)
- Lookup table/Histogram (LUT, HIST)
- Matrix Multiply Accelerator (MMA) – bolt on

C7x + MMA



Under the hood J721S2/J721E/J784S4

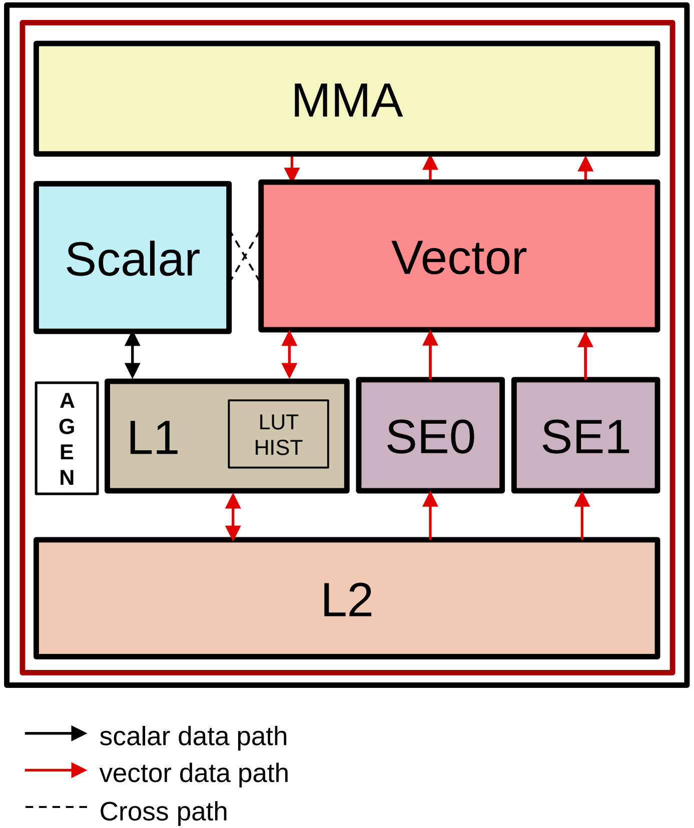
- **Data paths**
  - Scalar path – 64 bits
  - Vector path – 512 bits
  - Cross path – 64 bits
- C7x Load-Store to L1
  - 64 bits load || 64 bits store
  - 64 bits load || 512 bits store
  - 512 bits load || 64 bits store
  - 512 bits load || 512 bits store
- C7x Load using SE from L2
  - “Read Only” 2 x 512 bits
- C7x transfer to MMA (2 x 512 bits)
- C7x transfer from MMA (1 x 512 bits)



Under the hood AM62A

- **Data paths**
  - Scalar path – 64 bits
  - Vector path – 256 bits
  - Cross path – 64 bits
- C7x Load-Store to L1
  - 32 bits load || 32 bits store
  - 32 bits load || 256 bits store
  - 256 bits load || 32 bits store
  - 256 bits load || 256 bits store
- C7x Load using SE from L2
  - “Read Only” 2 x 256 bits
- C7x transfer to MMA (2 x 256 bits)
- C7x transfer from MMA (1 x 256 bits)

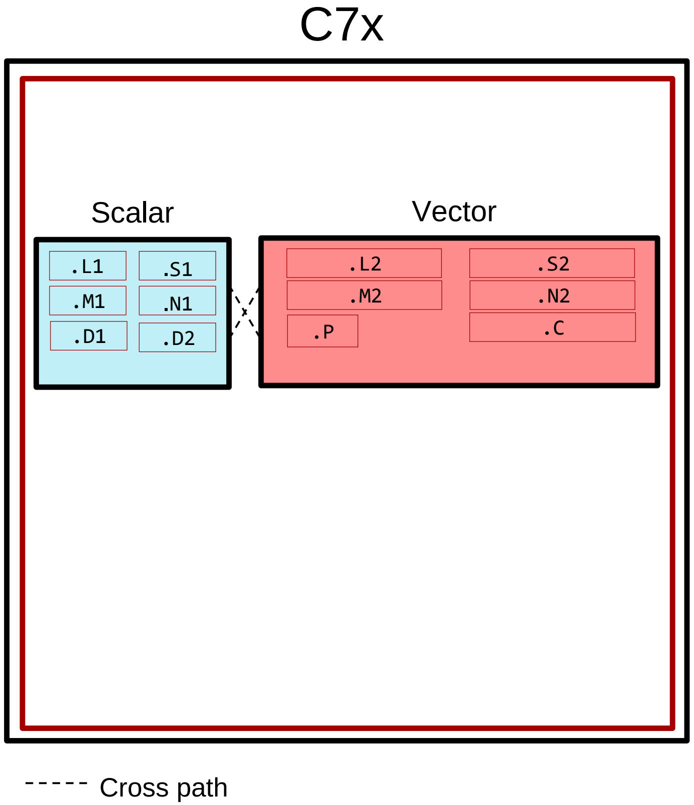
C7x + MMA



Functional Units

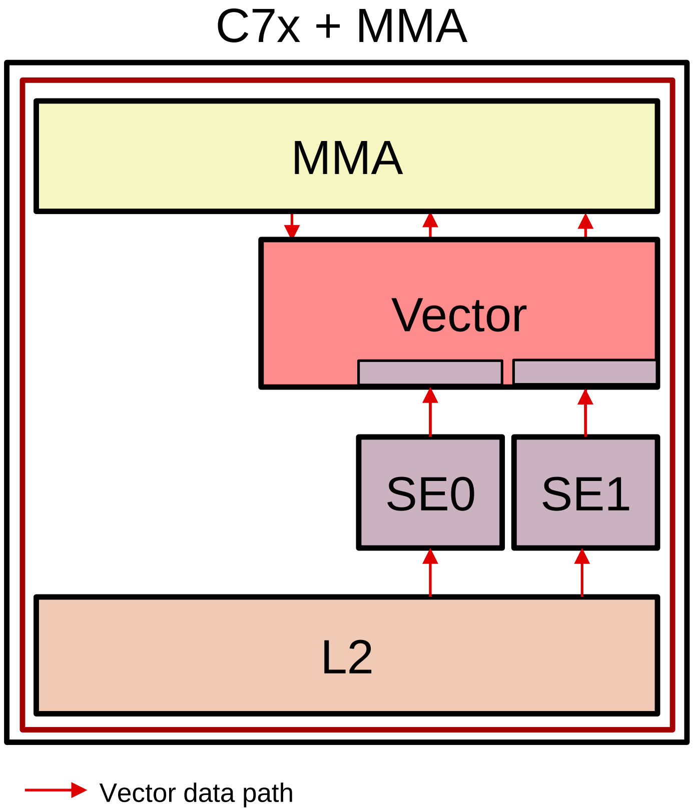
- **Functional Units (13)**
  - .L1/.L2 – Add/Sub/Move/Logical/Bitwise/Shift
  - .S1/.S2 – Add/Sub/Move/Logical/Bitwise/Shift
  - .M1/.M2 – Add/Sub/Multiply
  - .N1/.N2 – Multiply
  - .D1/.D2 –Load/Store/LUT/HIST
  - .C – Add/Permute/DOTP/SAD
  - .P – Vector predication
  - .B – Branch Predictor

Operation	Performance
16b fixed point MAC	128 MAC/cycle
32b multiply	32 multiply/cycle
32b Float ops	80 ops/cycle
8bit SAD	512 sad/cycle



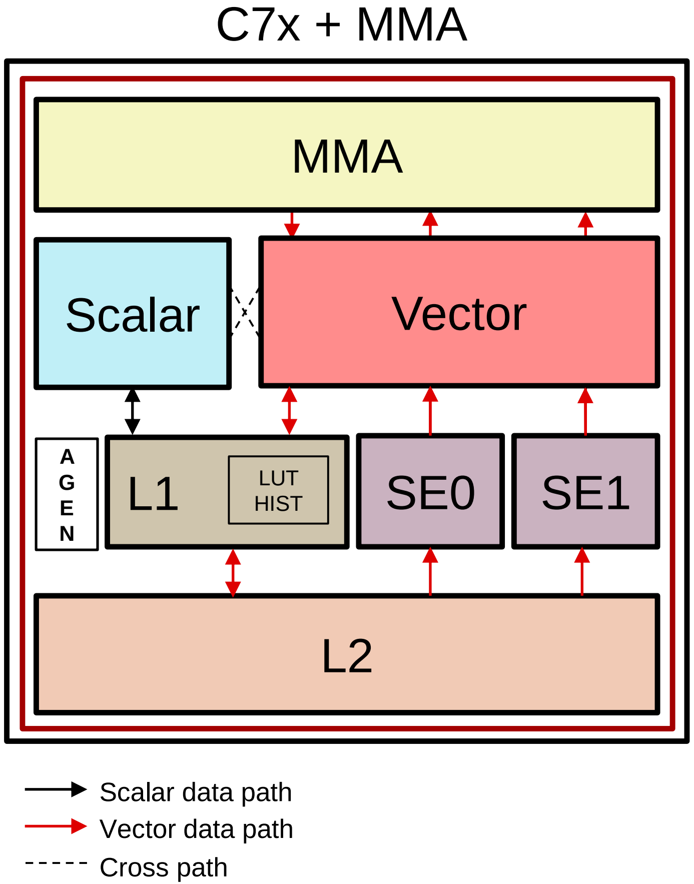
Streaming Engines

- **Streaming Engine**
  - Data forwarding engine not transfer engine
    - Forwards data from L2 and beyond directly to CPU boundary (C7x) or MMA memories (A and B)
  - Data formatting engine
    - Supports element promotion, decimation, duplication, transpose loads, predication
  - Provides 6D addressing
    - Access patterns up-to 6D can be programmed ahead.
    - 6D data is presented as (512bit for J721S2/J721E/J784S4 and 256bits for AM62A) vector per cycle.
  - Communicates with L2 memory controller for requests beyond L2 (L3, DDR)
  - Coherent with L1D data at stream open/close boundaries.
  - It’s a “Read Only” engine which feeds only vector path
  - Local cache (2KB) for reduced traffic at L2



Memory Layout J721E

- **Memory**
  - Level 1 memory (L1) at CPU clock
    - Program memory (L1P) 32kb
    - Data memory (L1D) 48kb
      - Separate 16 x 512b entry victim cache for stores
    - 1024b data throughput, 16 x 64b banks
    - ECC mode SECDED
  - Level 2 memory (L2) at CPU clock
    - Unified memory 512kb
    - Supports 4 masters (L1, SE0, SE1, DMA)
    - 2048b data throughput, 4 x 512b banks with 2 virtual banks each
    - ECC mode SECDED
  - Cache modes
    - L1P – $32kb (max) no SRAM mode
    - L1D - $32kb (max), $8kb (min) remaining as SRAM
    - L2 - $64kb (min) to $512kb (max)



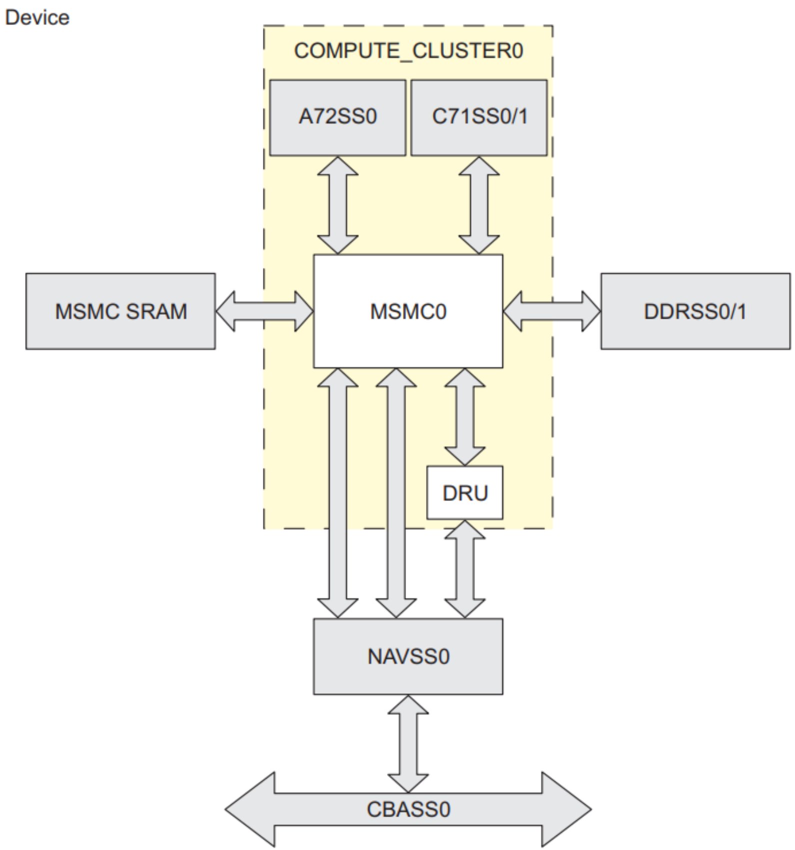
Under the hood AM62A

- C7x ISA and RISC-V G ISA extensions
  - Vector DSP, 40 GFLOPS
  - 256-bit vector width
  - Deep-learning Matrix Multiply Accelerator (MMA), up to 2 TOPS
- L1 memory architecture
  - 32KB I-cache
  - 64KB D-cache
- L2 memory architecture
  - Repurposing of its L2/EL2 embedded memory for use by SOC resources when C7x/MMA/DRU are disabled
  - 1.25MB L2 with ECC protection on L2 SRAM
  - Unified Memory Controller (UMC) facilitates L2 SRAM accesses from CPU and SOC (DMAs) as well as EMIF accesses from CPU.
- DRU (DMA engine) integrated that facilitates data transfer between L2, VPAC and EMIF
  - Data compression support
  - Event bus interface integrates with SOC DMSS
  - Tightly coupled with C7x/MMA (DRU is not available for use when the C7x/MMA is disabled)

MSMC Overview : J721S2

MSMC supports the following features

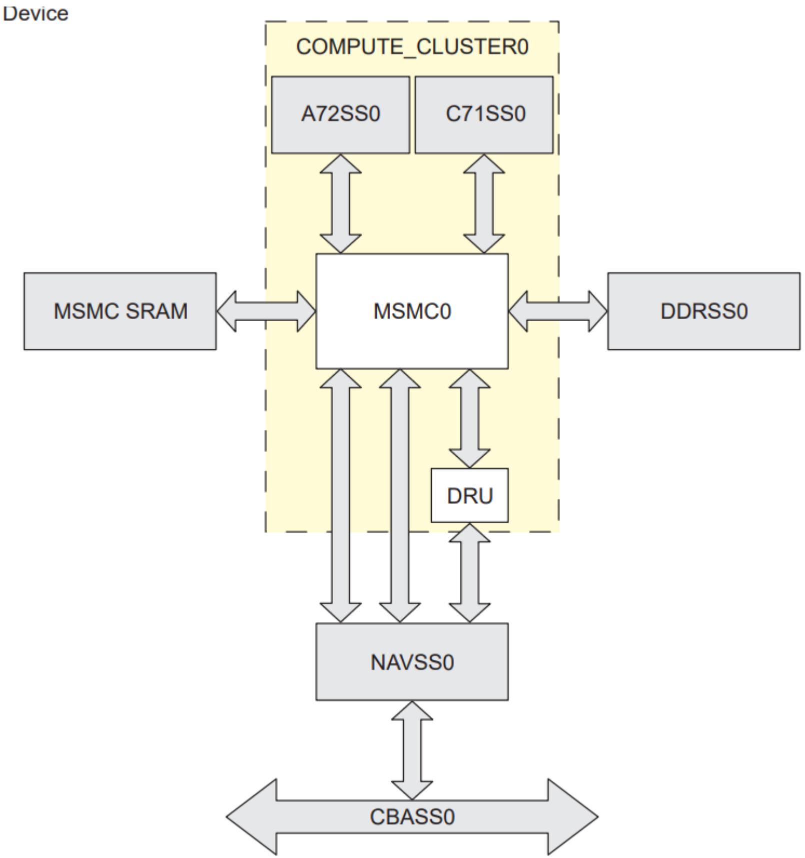
- 4MB (4 banks x 1MB) SRAM with ECC:
    - Shared coherent level 2/level 3 memory-mapped SRAM
    - Shared coherent level 3 cache
- 512-bit processor port bus and 40-bit physical address bus
- Coherent unified bi-directional interfaces to connect to processors or device masters
- One infrastructure master interface
- Dual external memory master interface
- Supports internal DMA engine – DRU (Data Routing Unit)
    - DMA in/out L2 SRAM, MSMC, DDR and system
    - L2, L3 cache pre-warming and post flushing
- Bandwidth management with starvation bound
- Two-level QoS support for real-time/nonreal-time split
- Security firewall flush support for SRAM/cache and external memory
- One interconnect messaging interface that supports DMA/prefetch requests to DRU



MSMC Overview : J721E

MSMC supports the following features:

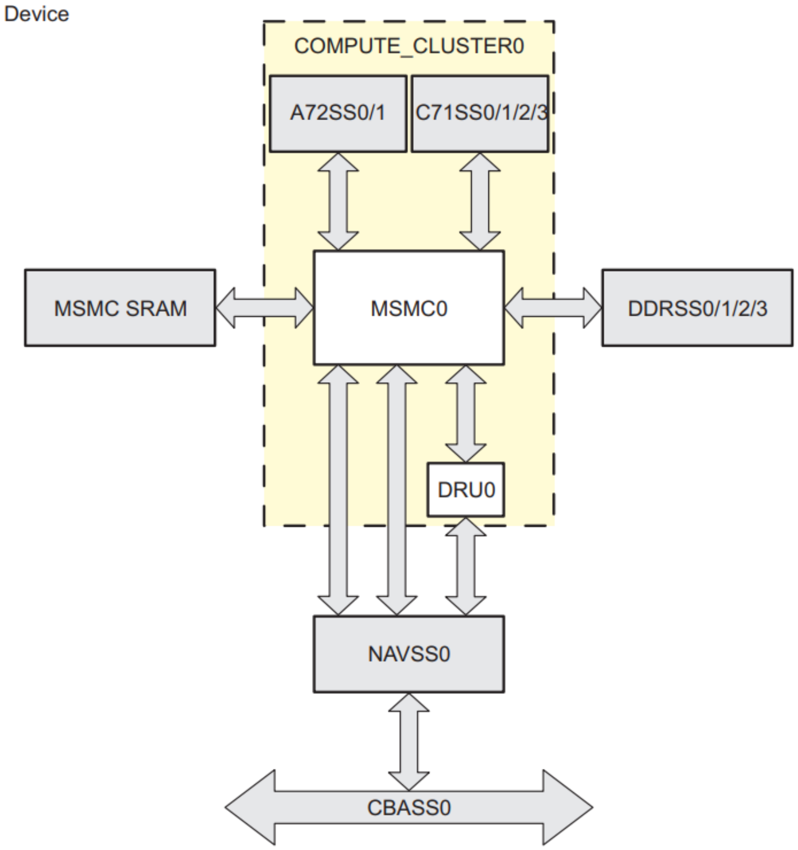
- 8MB (4 banks x 2MB) SRAM with ECC
    - Shared coherent level 2/level 3 memory-mapped SRAM
    - Shared coherent level 3 cache
- 512-bit processor port bus and 40-bit physical address bus
- Coherent unified bi-directional interfaces to connect to processors or device masters
- One infrastructure master interface
- Single external memory master interface
- Supports distributed virtual system
- Supports internal DMA engine – DRU (Data Routing Unit)
    - DMA in/out L2 SRAM, MSMC, DDR and system
    - L2, L3 cache pre-warming and post flushing
- Bandwidth management with starvation bound
- Two-level QoS support for real-time/nonreal-time split
- Security firewall flush support for SRAM/cache and external memory



MSMC Overview J784S4

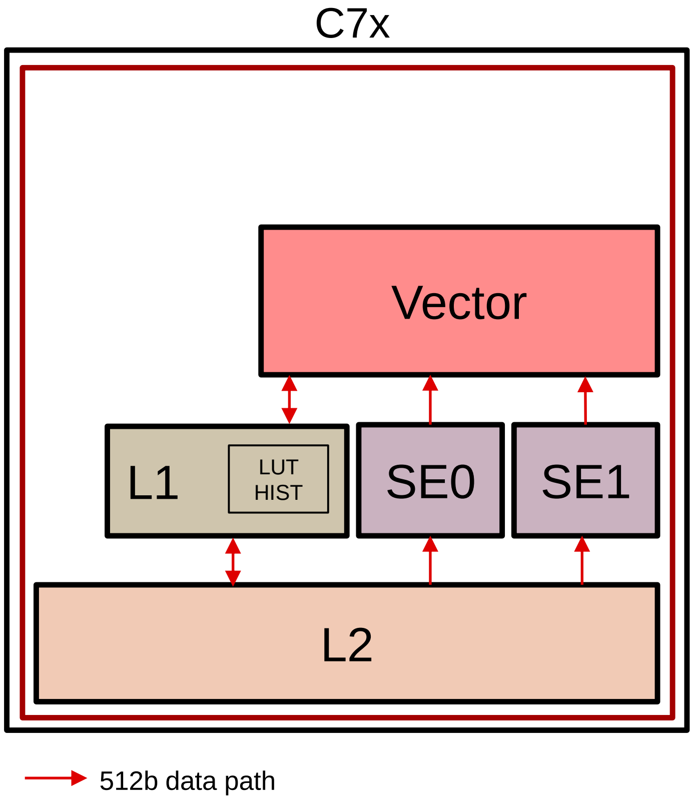
MSMC supports the following features

- 8MB (5 banks x 1.6MB) SRAM with ECC:
  - Shared coherent level 2/level 3 memory-mapped SRAM
  - Shared coherent level 3 cache
- 512-bit processor port bus and 40-bit physical address bus
- Coherent unified bi-directional interfaces to connect to processors or device masters
- One infrastructure master interface
- Single external memory master interface
- Supports distributed virtual system
- Supports internal DMA engine – DRU (Data Routing Unit)
  - DMA in/out L2 SRAM, MSMC, DDR and system
  - L2, L3 cache pre-warming and post flushing
- Bandwidth management with starvation bound
- Two-level QoS support for real-time/nonreal-time split
- Security firewall flush support for SRAM/cache and external memory



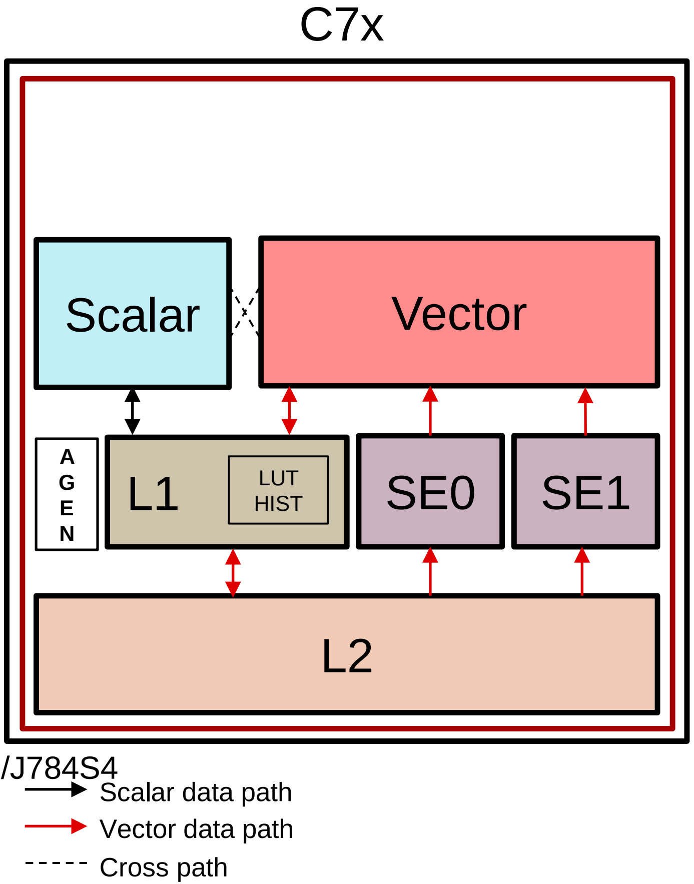
LUT and Histograms J721S2/J721E/J784S4

- **Lookup Table, Histogram**
  - Implemented in L1D, uses 16 x 64b banks
  - Data and Index supplied by registers or SE
  - Lookup table
    - Lookup in powers of 2 (1, 2, 4, 8, 16)
    - Supports - 8b, 16b, 32b
    - Table size (0.5KB, 1KB, 2KB, 4KB, 8KB, 16KB, 32KB)
    - Table size depends on number of ways
      - Eg. 16KB L1D SRAM split in 16 ways provides 1KB per way.
    - Number of bins depends on data type.
      - Eg. For 8b, 1KB table == 1024 bins
    - 1024 bit / cycle table initialization
  - Histogram
    - Supports 16 way histogram,
    - Supports – 8b, 16b, 32b
    - Supports weighted histogram

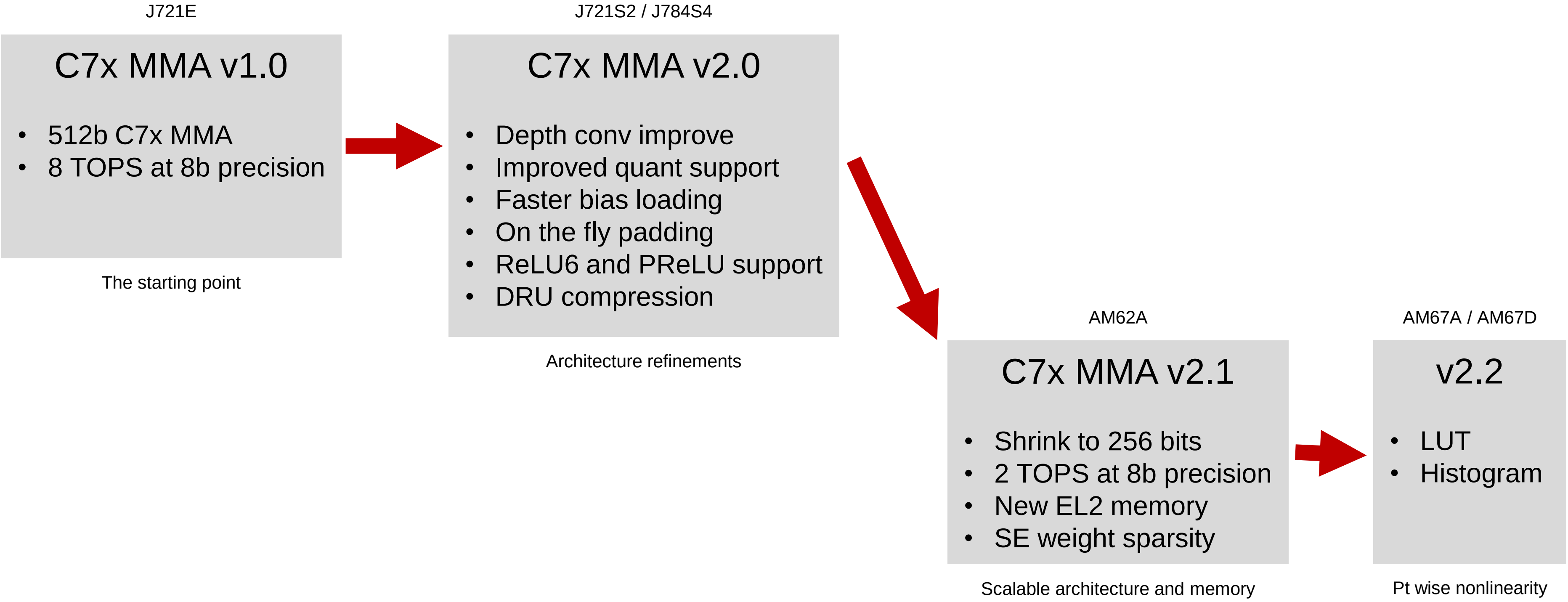


C7x DSP ISA

- **ISA**
  - Arithmetic, Shift, Logical instructions
  - Fixed point, Floating point, Complex type multipliers
  - Horizontal SIMD instructions, ADD, MIN/MAX
  - Byte permutation across SIMD lanes
  - Dedicated FIR instructions
    - FIR4 (4 tap), FIR8 (8 tap)
  - Dedicated DOTP instructions
    - DOTP2, DOTP4, DOTP8, DOTPMPN (flexible)
  - Sum of Absolute Differences (SAD)
    - 512 – 8bit, 256 16bit with stride support for J721S2/J721E/J784S4
    - 256 – 8bit, 128 16bit with stride support for AM62A
  - SORT16 instruction, ascending/descending
  - Galois Field multiply functions
  - WCDMA “Rake and Search” instructions
    - Up to 512 2-bit PN * 8-bit I/Q complex multiplies for J721S2/J721E/J784S4
    - Up to 256 2-bit PN * 8-bit I/Q complex multiplies for AM62A



C7x MMA From Past To Present



J721E

C7x MMA v1.0

- 512b C7x MMA
- 8 TOPS at 8b precision

The starting point

J721S2 / J784S4

C7x MMA v2.0

- Depth conv improve
- Improved quant support
- Faster bias loading
- On the fly padding
- ReLU6 and PReLU support
- DRU compression

Architecture refinements

AM62A

C7x MMA v2.1

- Shrink to 256 bits
- 2 TOPS at 8b precision
- New EL2 memory
- SE weight sparsity

Scalable architecture and memory

AM67A / AM67D

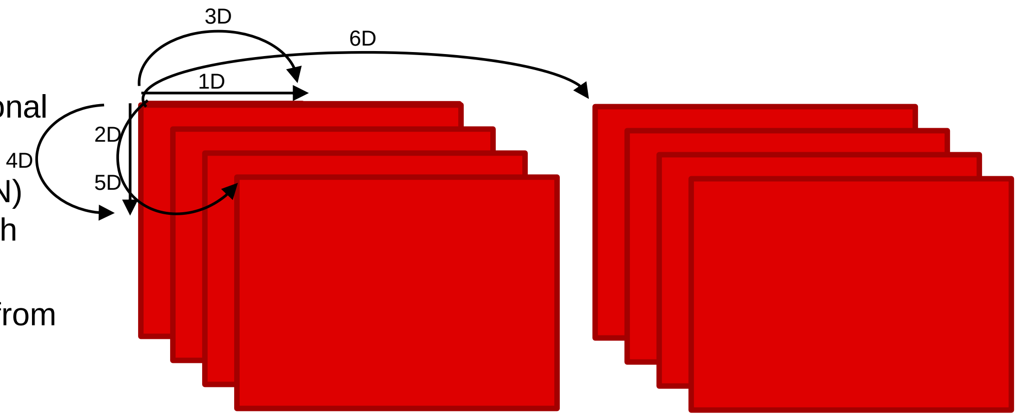
v2.2

- LUT
- Histogram

Pt wise nonlinearity

Under the hood

- **Address generators (AGEN)**
  - Computer vision applications have multi-dimention access patterns which take up registers and functional units
  - C7x core has dedicated Address GENerator (AGEN) unit which computes multi dimensional offsets which can be used with regular load/store instructions
  - Up to 4 AGEN units supported in each core. Apart from 2 in Streaming Engine.
  - Supports up to 6D addressing



```
offset = ICNT0 + ICNT1 * DIM1 + ICNT2 * DIM2 + …
val = pSrc[offset]
```

C66x vs C7x - Summary

	C66x DSP	C7x DSP J721S2/J721E/J784S4)
DSPType	True 32 bit 32bit/64bit floating point types 6bit/32bit complex types	True 64 bit 32bit/64bit floating point types 6bit/32bit complex types
Functional Units	8 functional units (.L1/.L2, .S1/.S2, .M1/.M2, .D1/.D2)	12 functional units (.L1/.L2, .S1/.S2, .M1/.M2, .N1/.N2 .D1/.D2, .C, .P)
Data paths	2 x 64 bit , 64 bit cross path	64 bit + 512 bit + 2x512 (read only), 64 bit cross path
Registers	32x2 – 32 bit registers	16-64 bit global, 24-64 bit local, 16-512 bit global, 24-512 bit local, 8-64 bit local (.P)
Cache	32KB L1P + 32KB L1D, 256KB L2	32KB L1P + 32KB L1D + 16KB L1SRAM, 512KB L2
Multipliers	32 -16bit fixed, 8 – 32bit fixed / floating	128 - 16bit fixed, 32 – 32bit fixed / floating
Operations	32-GMAC, 16-GFLOPS at 1 GHz	128-GMAC*, 80-GFLOPS at 1 GHz
Transfer engines	IDMA (2 channels),1D – 32bit	Streaming Engines (2 sets), 6D – 512bit, read-only
Coherency	Coherent with L2	Fully coherent with L2, L3, DDR
Safety	ECC – SED, L1	ECC – SECDED, L1/L2, SE FIFO
SIMD	Packed SIMD (8b, 16, 32b, 64bit) Inter or Vertical SIMD (.L, .S, .M, .D)	Packed SIMD (8b, 16, 32b, 64bit) Inter or Vertical SIMD (.L, .S, .M, .N, .C, .D) Intra or Horizontal SIMD (.C)
HW Acceleration	SPLOOP HW	NLC (Nested Loop Controller), branch predictor Lookup Table, Histogram

*Excluding MMA mac