Guiding Power/Quality Exploration for Communication-Intense Stream Processing

Hamed Tabkhi, Majid Sabbagh, Gunar Schirner

Department of Electrical and Computer Engineering
Northeastern University

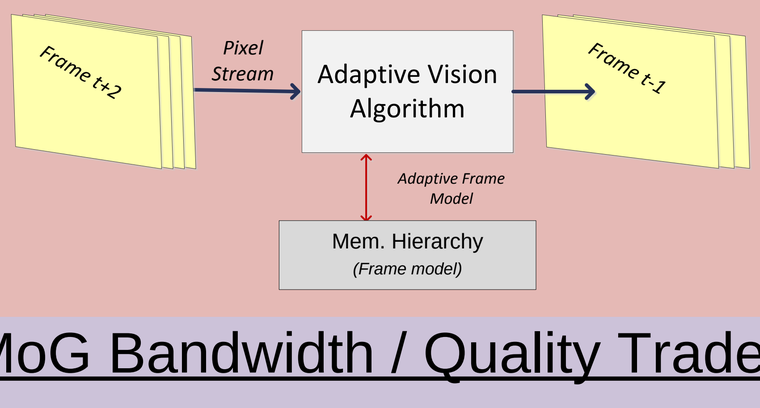
Northeastern

Embedded Systems Laboratory

Adaptive Stream Processing

1) Adaptive Vision Streaming

- Often based on machine-learning algorithms
  - e.g., Mixture of Gaussians (MoG), Support Vector Machine (SVM)
- Realize complex vision processing with high quality
  - e.g., object detection, tracking, classification
- Continuously update model of frame characteristics
  - Frame model is often large (e.g. MoG with 248 MB per frame 1080p)



Mixture of Gaussian (MoG)

1) Mixture of Gaussian (MoG)

- Extracts ForeGround (FG) pixels from BackGround (BG) scene
- Adaptive learning-based BG tracking for static camera position

2) MoG Resource Demands

- Computation:
  - 24.3 GOPs at 1080p60 (SW infeasible)
  - 20 (float) or 13 (integer) Blackfin DSP)
- **Communication:**
  - 32 bits per Gaussian parameters (*weight, mean, standard deviation)*
  - Saturating LPDDR2

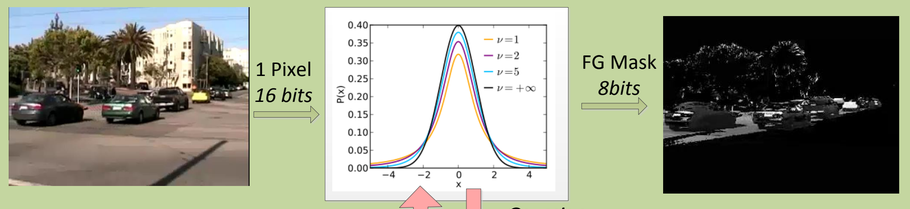
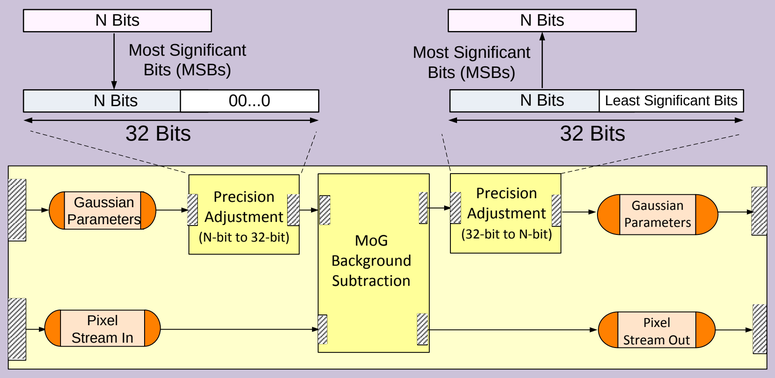


Image size	GOPs	Blackfin cores	Bandwidth [MB/Sec]	LPDDR2 Utili.
1920*1080	24.3	13	7440	Saturated
1280*960	14.4	8	4380	70%

MoG Bandwidth / Quality Trade-Off

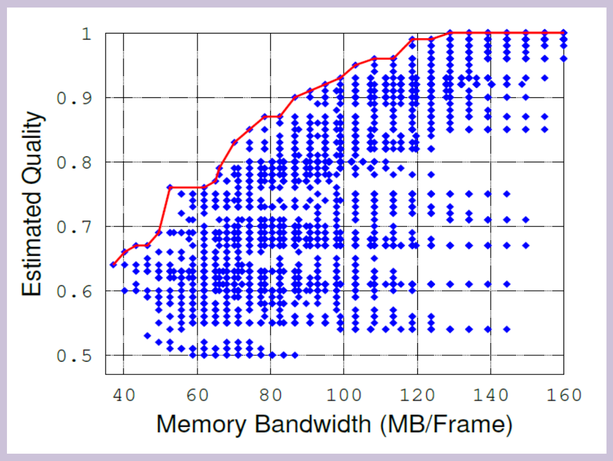
1) Bandwidth Reduction

- Reduce precision of frame model
  - At cost of some quality
- MoG Frame Model: Gaussian parameters
  - Transfer/store only most significant bits
    - E.g. 13bit instead of 32bits
    - More elaborate bit allocations possible
  - Realized by precision adjustment blocks
    - Streamed access hides latency
- Tradeoff bandwidth v.s. Quality



2) Trade-off Analysis

- Quality assessment use MS-SSIM [Wang et. al 2003]
  - More expressive than PSNR
- Quality increases with BW
  - No significant improvement after 130 MB/Frame
  - The Pareto front (red line)
    - E.g., 95% quality at 63% bandwidth

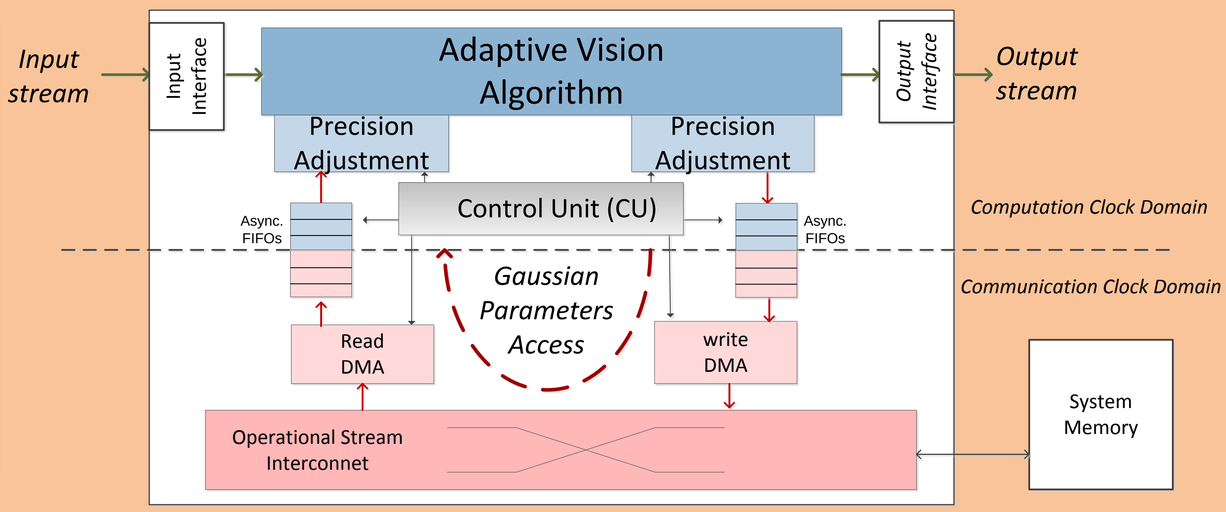


MoG Architecture Template

1) Communication Components

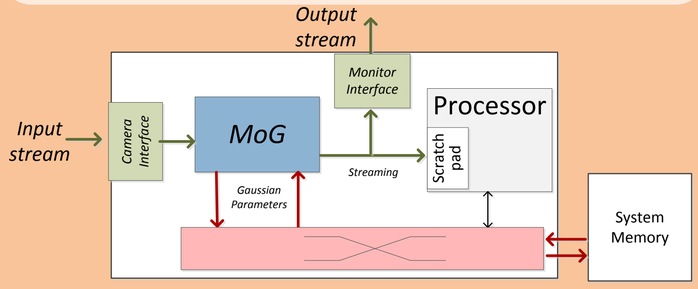
- Independent traffic management
  - Separate clock domains for computation and communication
- Dedicated precision adjustment blocks
  - Transferring most significant bits
- 2 DMA channels for Gaussian parameters
  - Connected to AXI with burst transfer
- Async FIFOs
  - Bridge clock domains
  - Compensate for slow interconnect

  (148.5Mhz pixel v.s. 125MHz AXI fabric)



2) System Integration

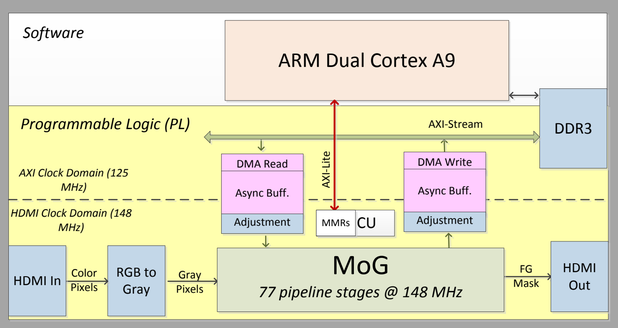
- Direct access to system I/O interfaces
- Keep the streaming data on-chip
  - Remove costly memory interaction
- Enable peer-processor arrangement
  - Host processor only performs first initialization

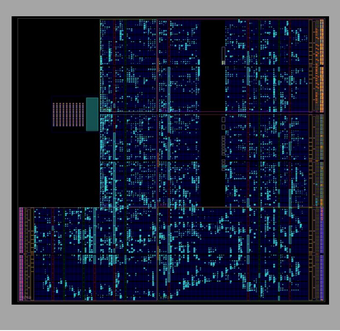


Experimental Results

1) Zynq 7020 Realization

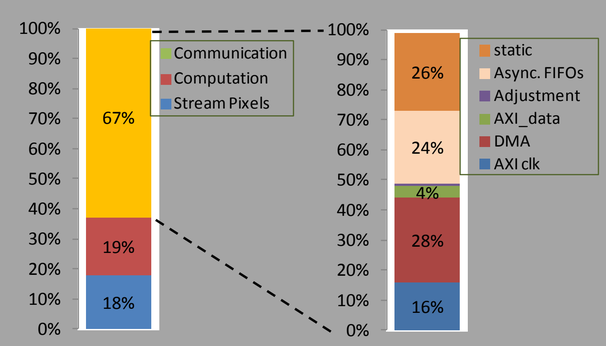
- Design spreads over chip
- Significant routing overhead
- 34% DSP slice utilization
- Resolution: 1080p @ 30FPs
  - Limit: peak memory bandwidth of 4.2 GBs





2) Power Consumption

- 600x more power efficient than SW (Cortex A9)
- 480mWatt on-chip power
  - Only 19% for computation
  - Only 1% for precision adjustment block
  - 67% for transferring Gaussian parameters
  - 28% and 26%, for DMA and Async. FIFOs



3) Power/Quality Trade-off

- Quality requirement results in:
  - Bandwidth requirement
  - Power consumption
- E.g. 100% quality -> 380mW
- 1/3rd power of [Appiah 2005]