National Taiwan University

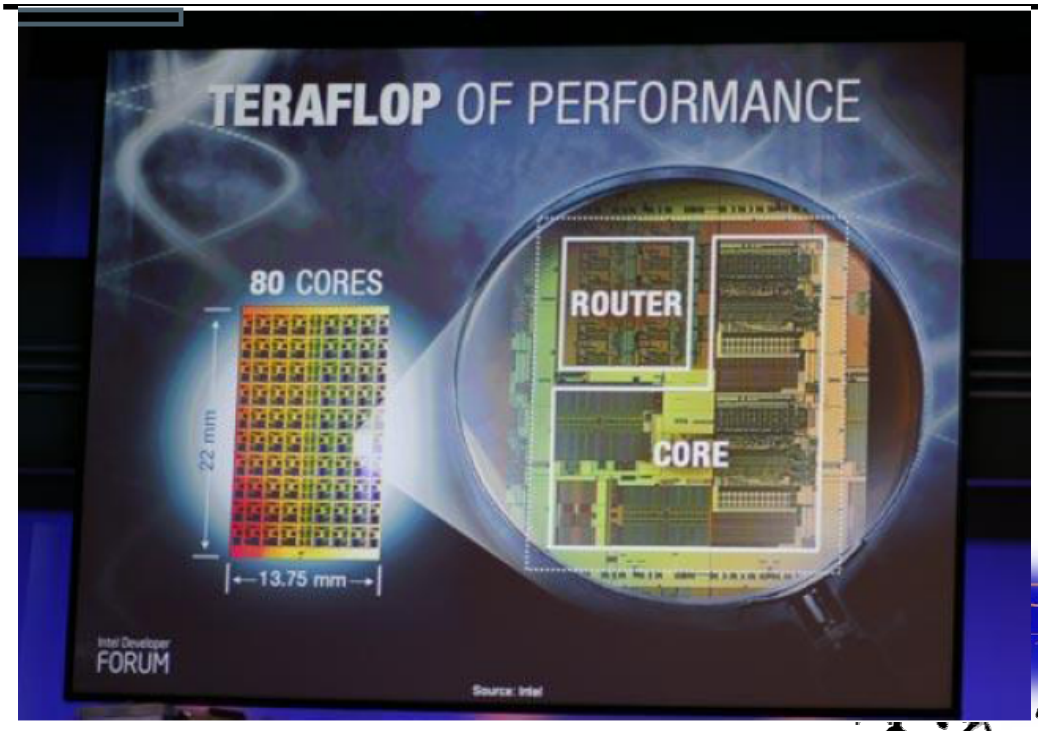
Scalable and Reconfigurable Stream Processor for Mobile Multimedia Systems

Liang-Gee Chen

Distinguished Professor
General Director, SOC Center
National Taiwan University

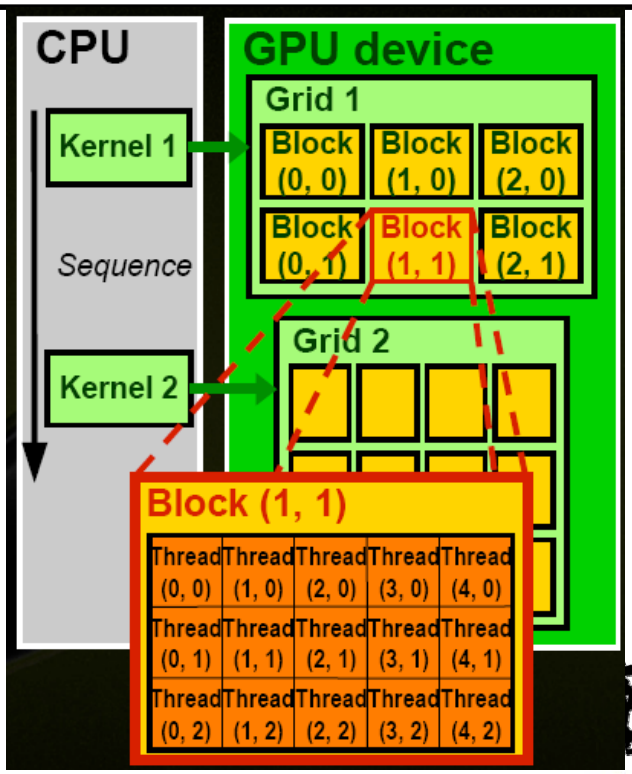
DSP/IC Design Lab, GIEE, NTU

Multicore SoC is coming



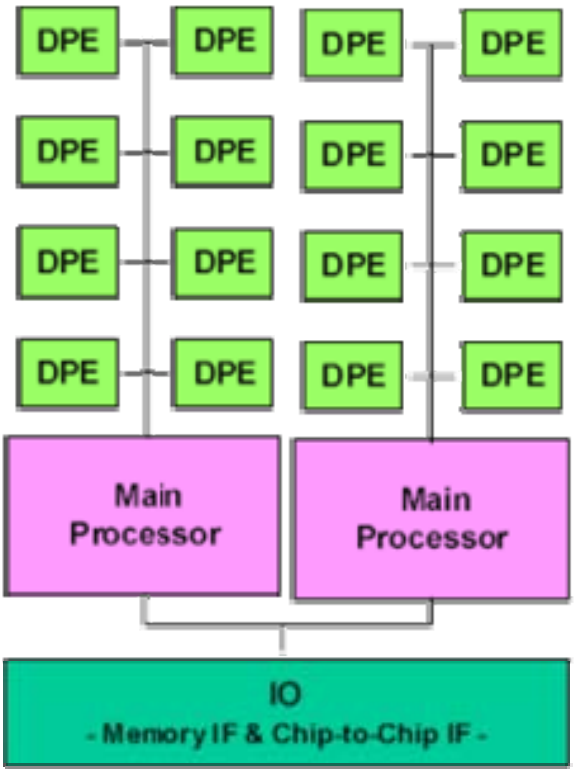
Source: 2007 ISSCC and IDF

Multithread Multiprocessor SM Core



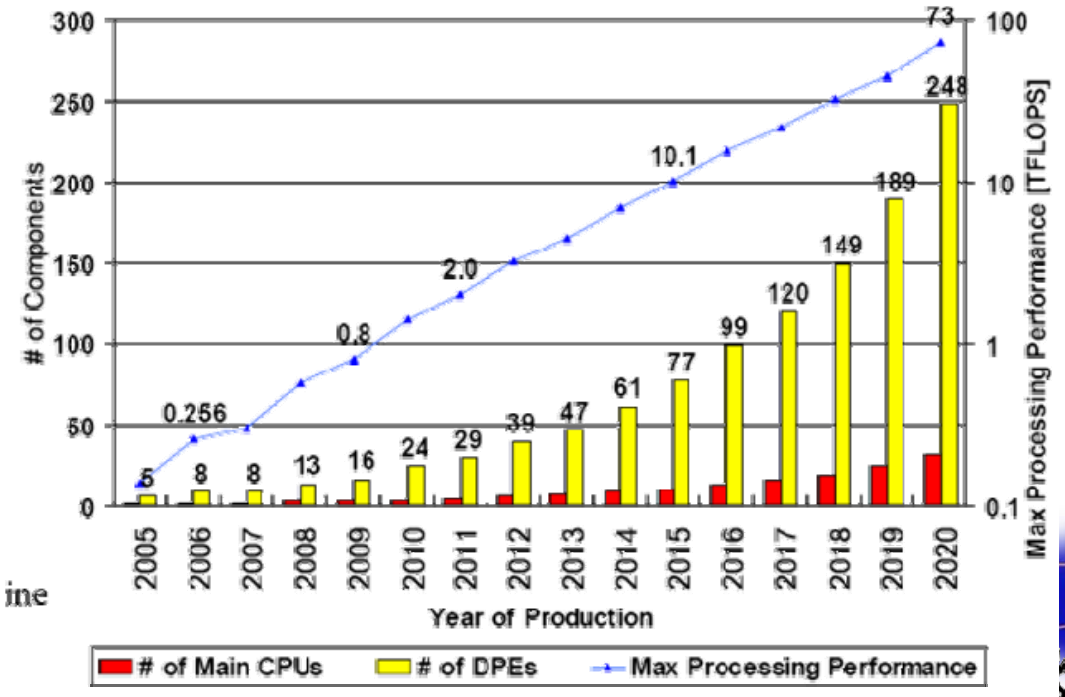
Source: nVidia

Typical Architecture of MPSOC



Source: ITRS

The Perspective for MPSoC



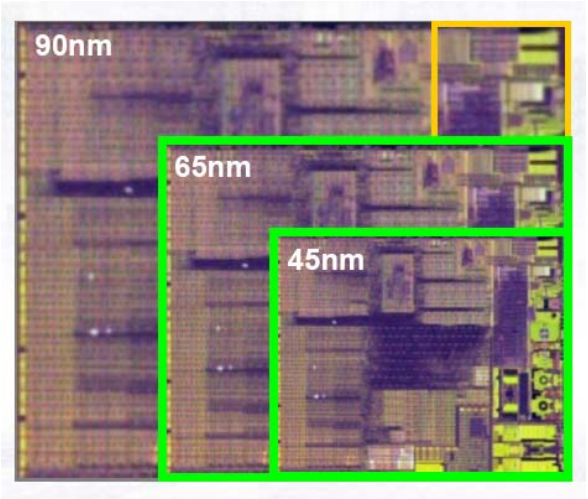
Source: ITRS, IEK/ITRI 2007

National Taiwan University

Driving Forces for MPSOC

- Technology Push: Semiconductor technology
  - Moore's law is still working

"Number of transistors on a chip doubling every 18-24 Months"



- Same design with
  - Smaller area
  - Lower cost
- Same area with
  - more functionality

National Taiwan University

Driving Forces for MPSOC

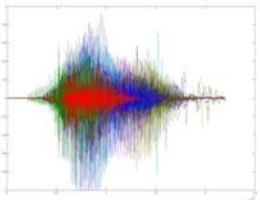
- Demand Pull
  - RMS (Recognition, Mining and Synthesis) for Server
  - MMM (Mobile MultiMedia) for end products

Mobile multimedia platform

- Mobile device becomes the personal multimedia platform

Multimedia Signal Processing

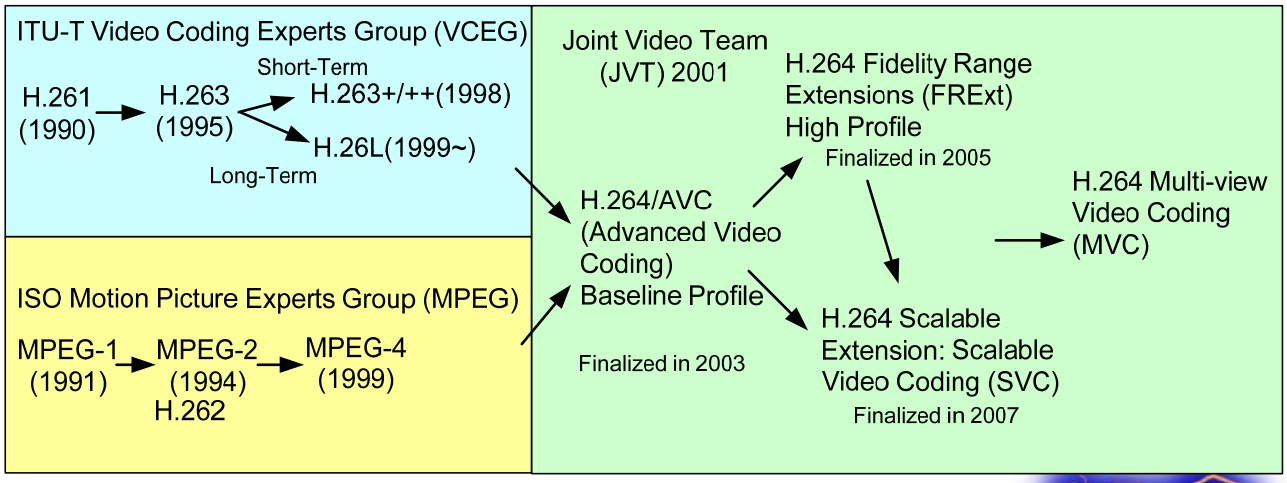
- **Signals:** Image, Speech, Audio, Graphic
- **Representations:** Bitstream, Frame, Pixel
- **Applications:** 2D/3D Modeling, Animation, Content Based Indexing and Retrieval, Storage and Transmission
- **Key Technology:** Video Coding and Graphics





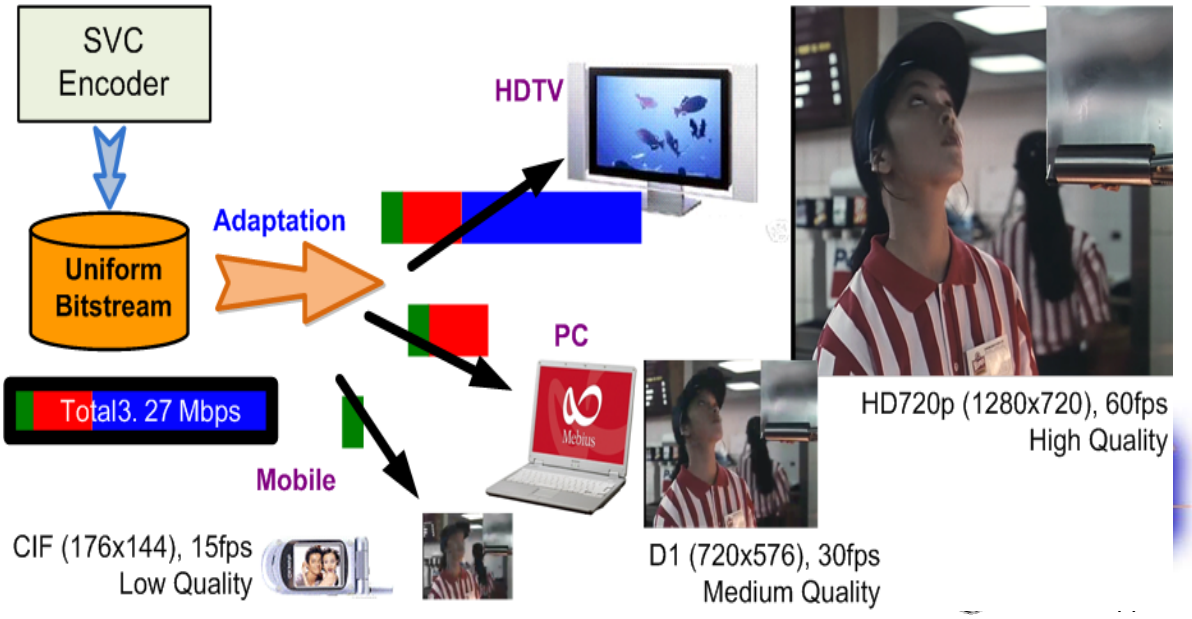
Ref: Prof. Pirsch Keynote on ICME 2007

Evaluation of Video Coding Standards



Services with Scalable Functions

- From compression efficiency to specific functionalities



Evolution of Coding Tools

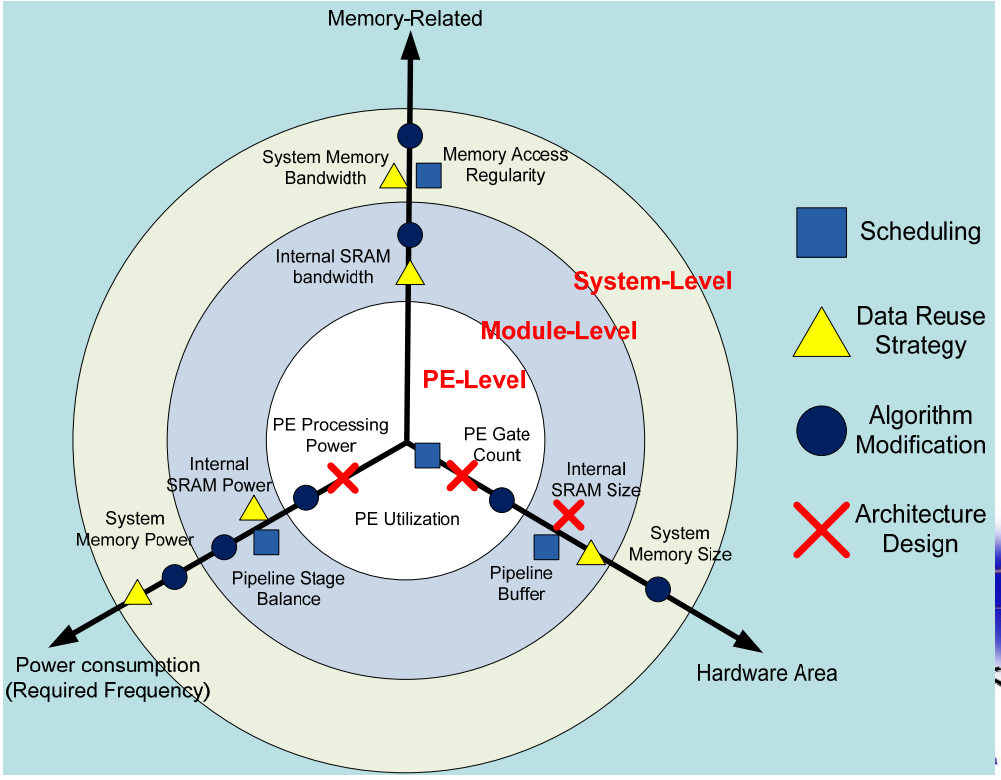
- New adaptive prediction modes, function-oriented tools

Tool	MPEG-2	MPEG-4 ASP	VC-1	H.264 Baseline	H.264 High profile	H.264 SVC
Frame Type	I, P, B	I, P, B	I, P, B	I, P	I, P, B	I, P, B, EI, EP, EB
Macro-block partition in MC	16x16	16x16,8x8	16x16,8x8	16x16…4x4 (7 modes)	16x16…4x4 (7 modes)	16x16…4x4 (7 modes)
Motion-vector precision	1/2	1/4	1/4	1/4	1/4	1/4
Intra-prediction modes		2	2	13	22	22
Transform	DCT	DCT	DCT*	4x4 Integer*	Integer*	4x4/8x8 Integer*
Loop filter			V	V	V	V
Entropy coding	VLC	VLC	CAVLC	CAVLC	CAVLC, CABAC	CAVLC, CABAC
Special features		global motion compensation		multiple reference frames	8x8 transform/prediction	Hierarchical B-frame, inter-layer prediction, CGS/MGS/FGS

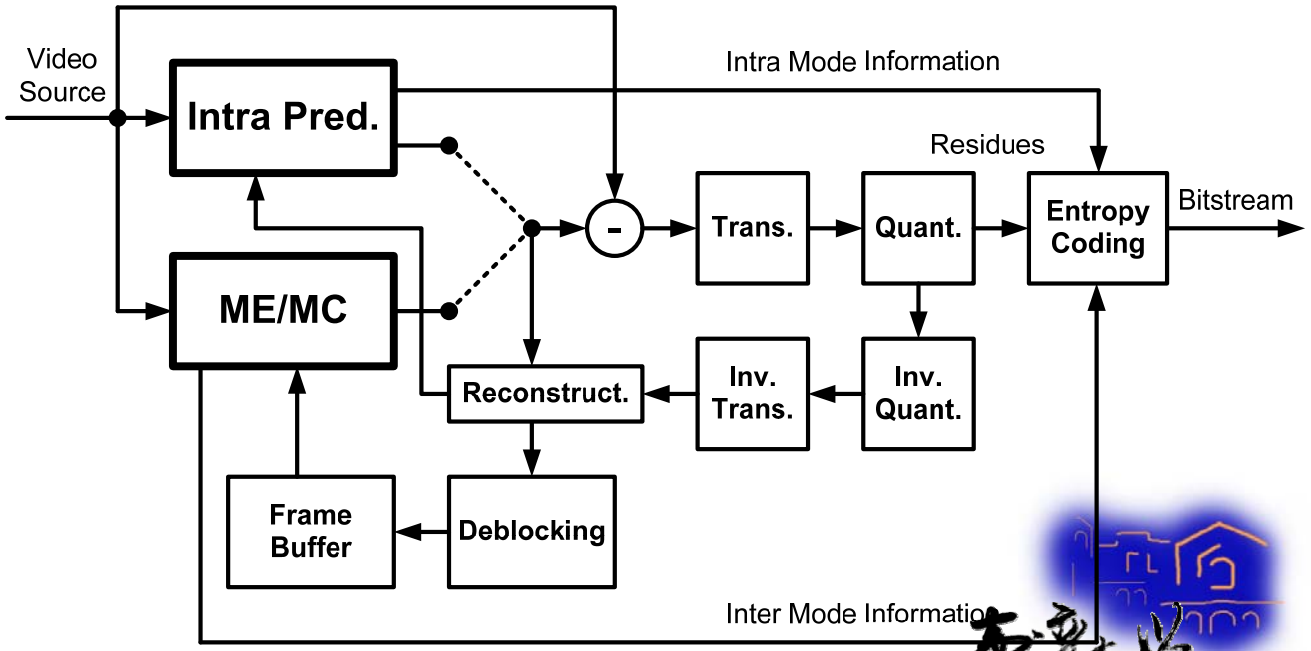
Architectural Perspective for Multimedia Processing

- Parallel Processing
  - Requirement: Processing of independent segments
  - Pipelining, systolic processing
  - Task level parallelism
  - Instruction level parallelism
  - Data level parallelism
- Stream Processor
  - Requirement: Processing of dependent segments
  - Scalable architecture
  - Multi-thread and cache

Design Space in Various Levels

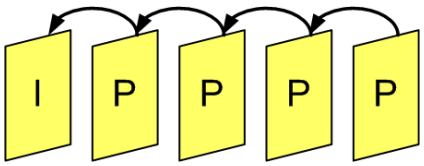


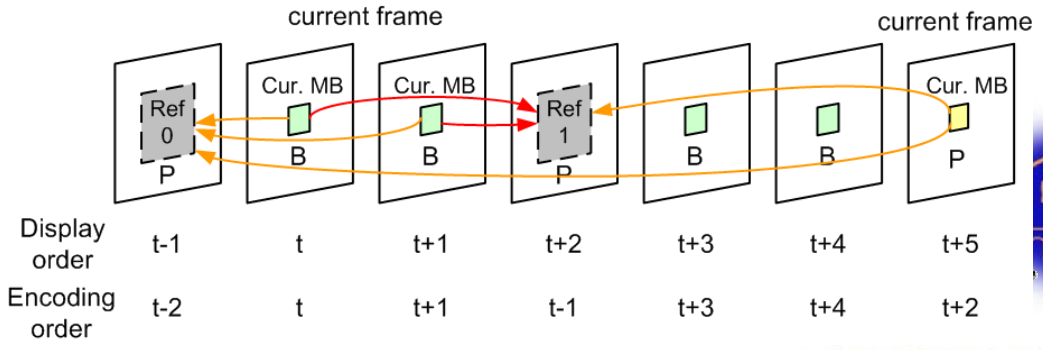
H.264 Encoding as Example



Data Dependency Analysis

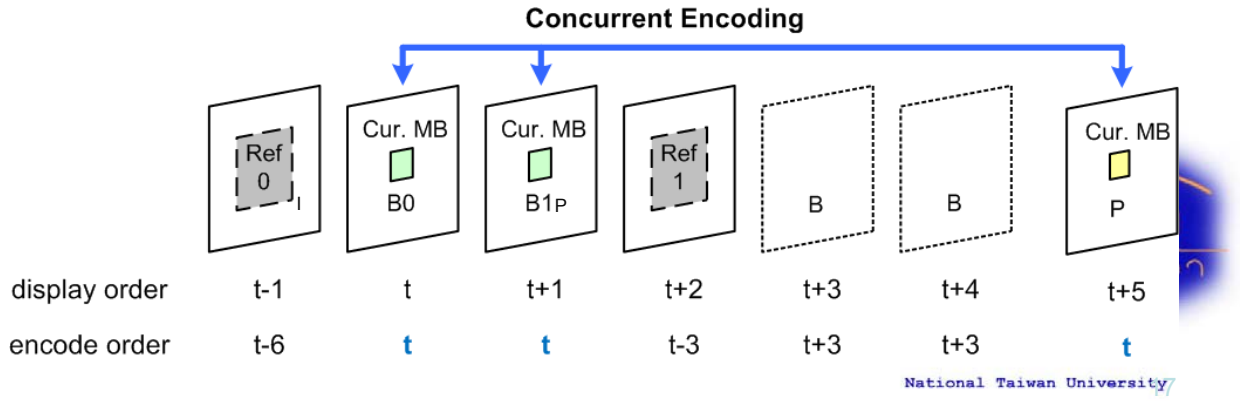
- P-frame scheme
  - Referenced frames are all different
- B-frame scheme
  - The B-frames may have the same referenced P-frames
  - The next P-frame also has the same reference frame





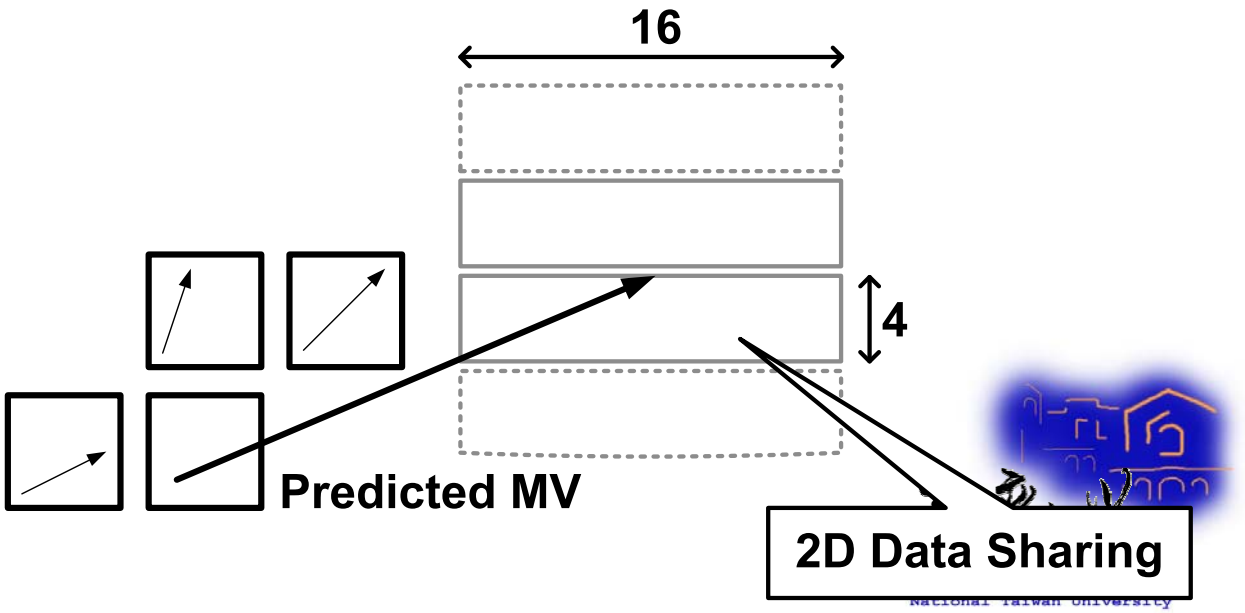
Frame-Parallel Encoding Scheme

- Encode frames of same reference frames in parallel
  - The MBs of same location are encoded simultaneously
  - Achieve frame-level data reuse
  - For IBBP scheme, 66% system memory bandwidth is saved

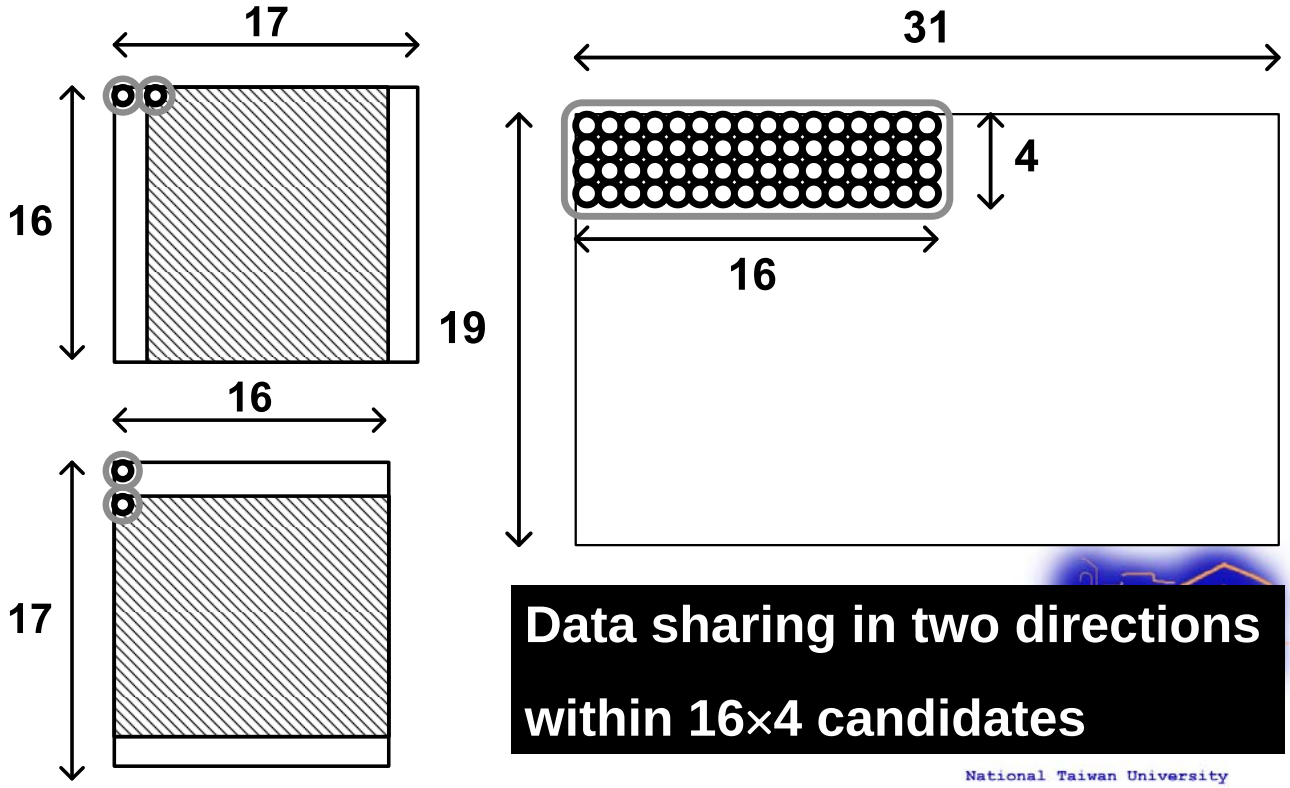


ME Algorithm

- **Predicted moving window (PMW) search**
  - Adaptive 16×16／16×8 window size



Concept of Data Sharing

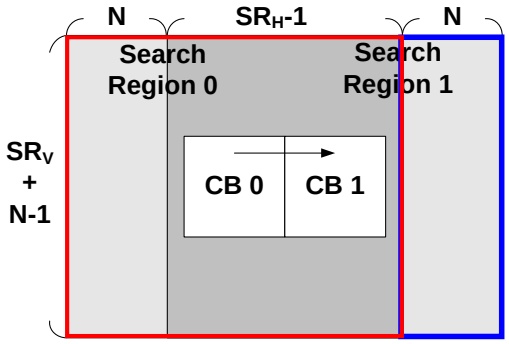


Level C and Level D Schemes

- Level C
  - Horizontal data reuse

$$Ea_{LevelC} \approx \frac{N \times (SR_V + N - 1)}{N \times N} \approx 1 + \frac{SR_V}{N}$$

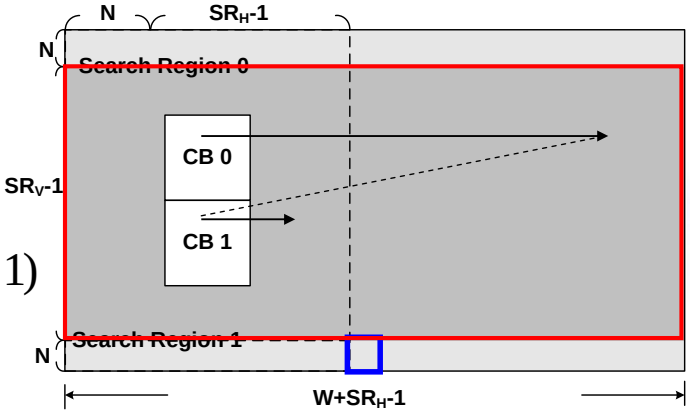
$$SRB = (SR_H + N - 1)(SR_V + N - 1)$$



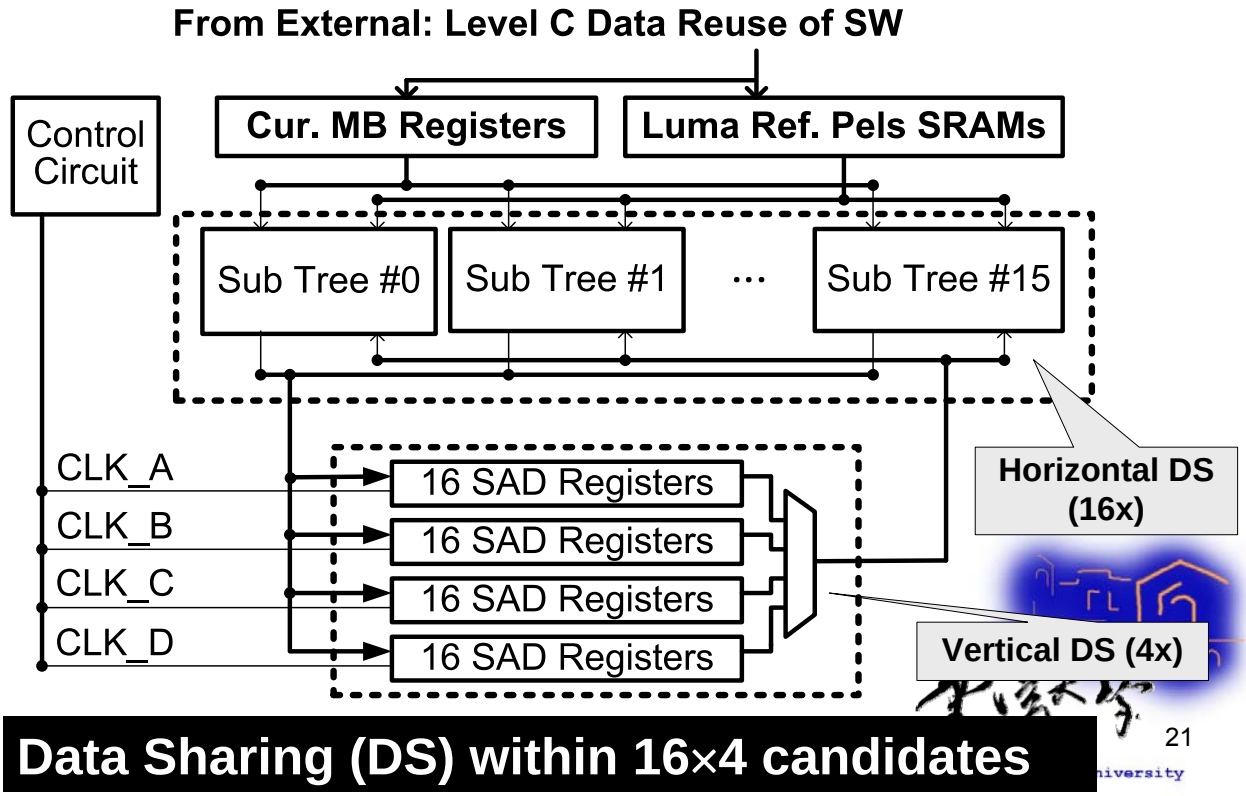
- Level D
  - Fully data reuse

$$Ea_{LevelD} = 1$$

$$SRB = (SR_H + W - 1)(SR_V - 1)$$

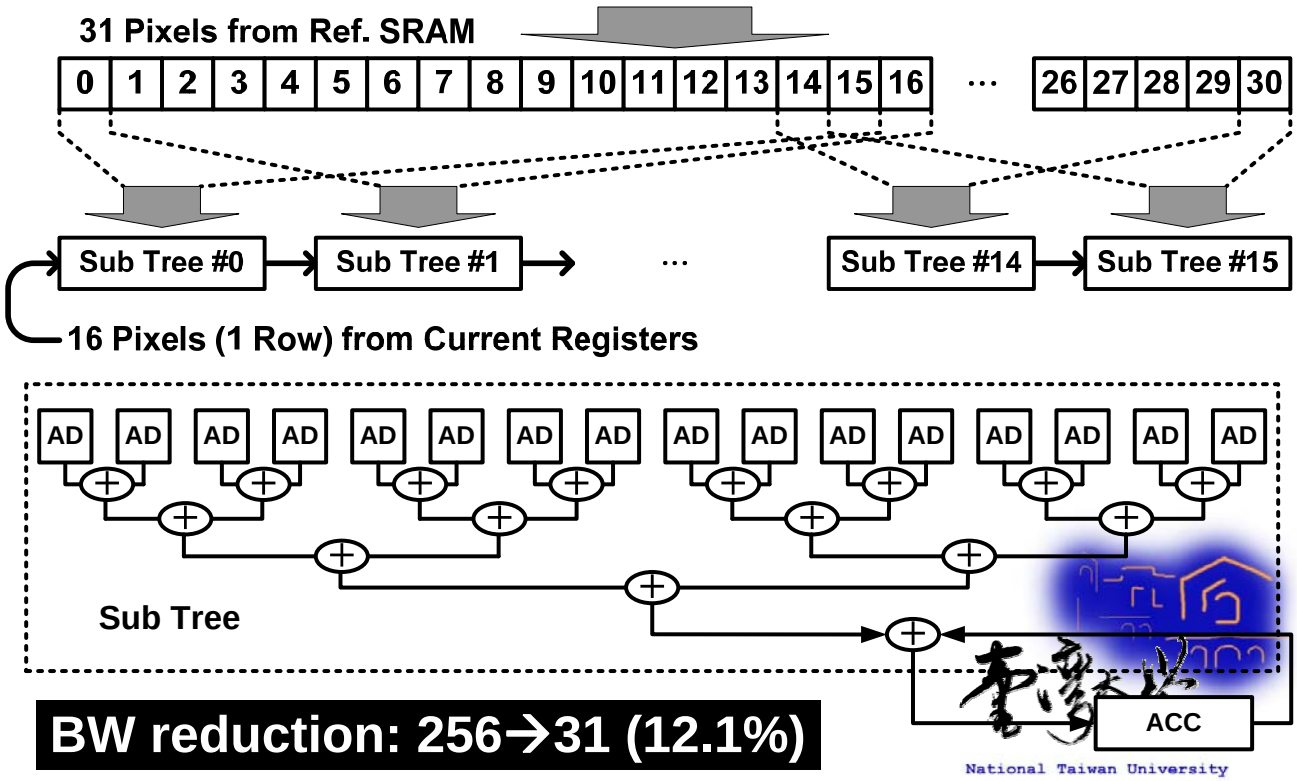


2D Bandwidth Sharing ME Architecture



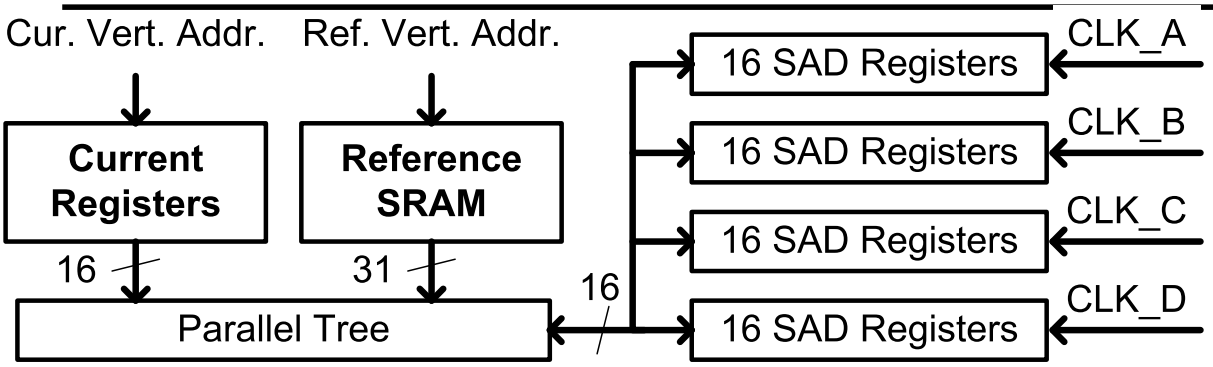
Data Sharing (DS) within 16×4 candidates

Horizontal Data Sharing

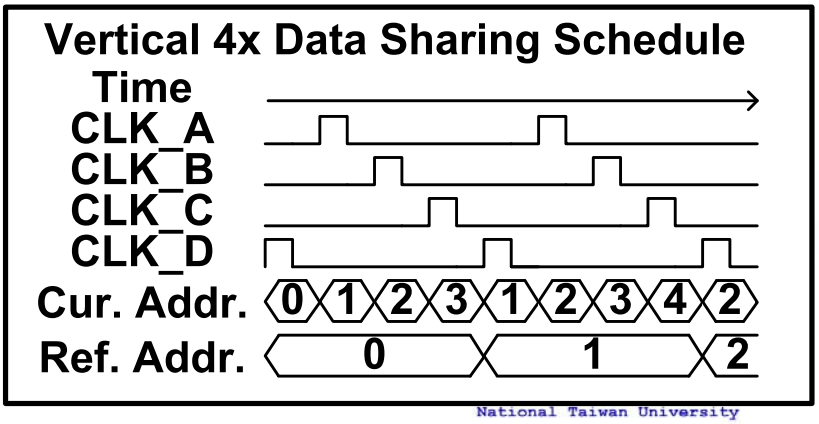


BW reduction: 256→31 (12.1%)

Vertical Data Sharing



BW reduction: 16x4→19 (29.7%)



Main Concepts of the FME Stage

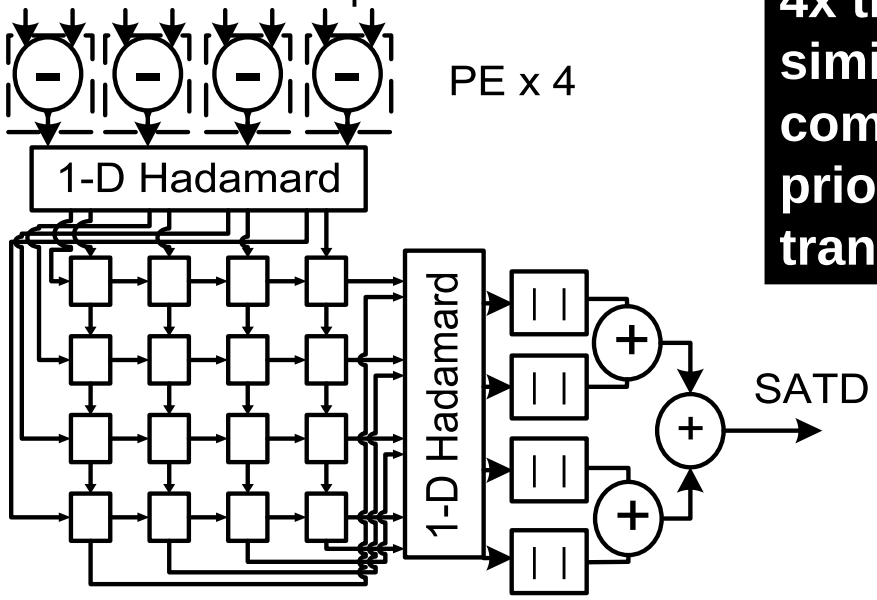
- **Thorough parallelization of each block with full pipeline and high utilization**
  - Sequential process of VBS
  - The least common factor of VBS is 4x4.
  - The SATD involves 4x4 Hadamard transform.

1. Design a 4x4-block processing unit (PU)

2. Apply the folding technique for larger blocks to iteratively utilize the PUs

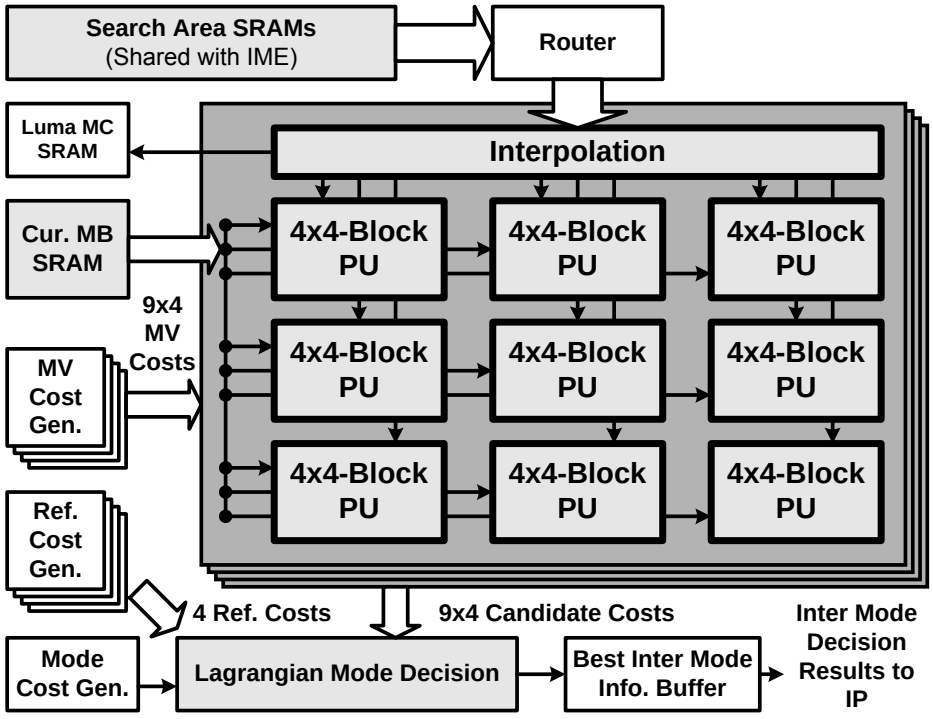
4x4-Block Processing Unit (PU)

A Row of Four Cur. MB Pels and
A Row of Four Interpolated Ref. Pels



4x throughput but similar gate count compared with the prior sequential transform design

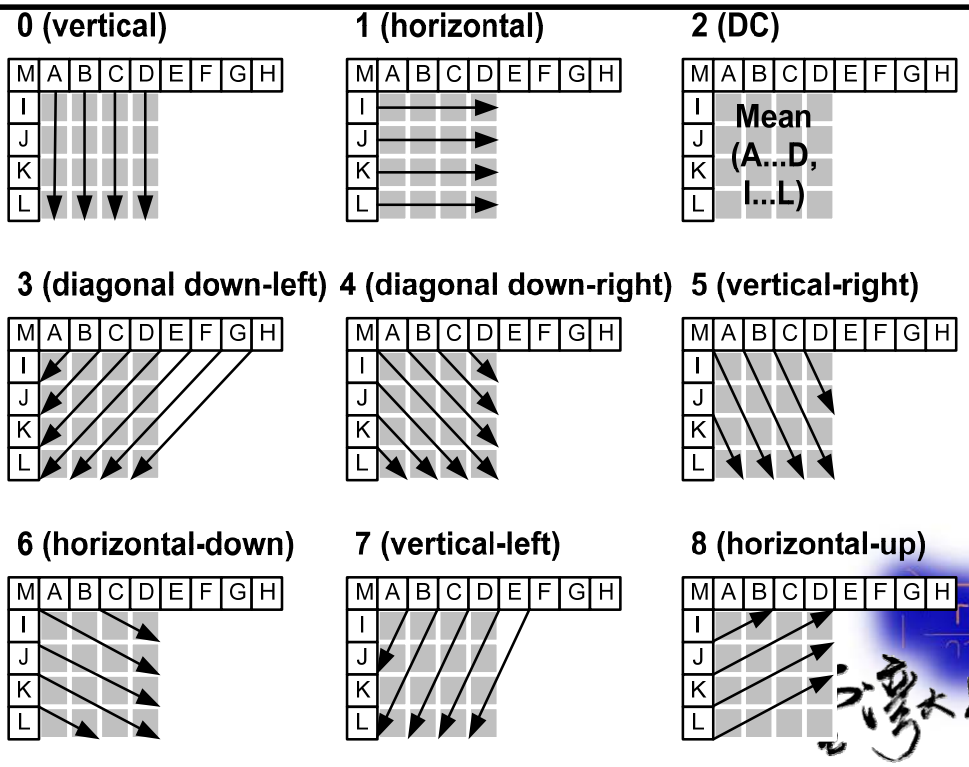
Parallel Configuration of PUs



Shared interpolation saves 764K logic gates.

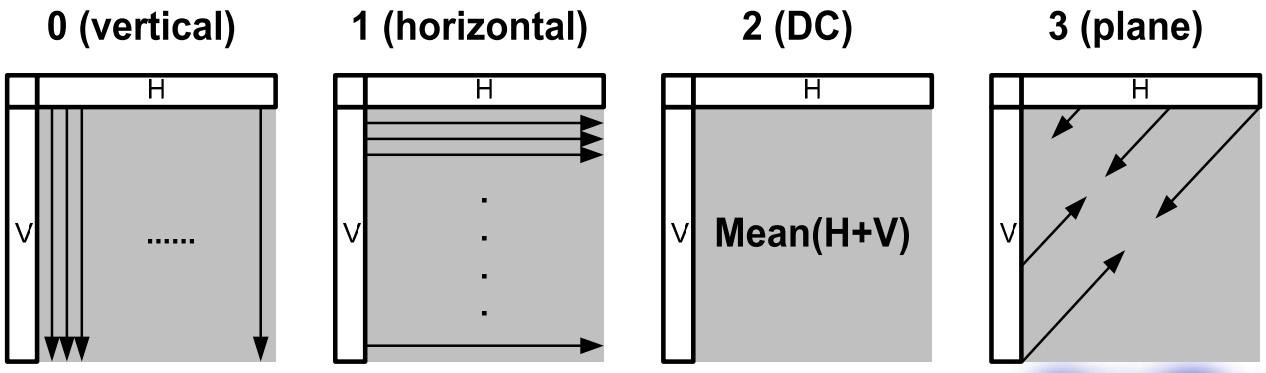
Nine PUs of Adjacent 3x3 Candidates for Every Ref. Frame

4x4 Intra Prediction (I4MB)



National Taiwan University

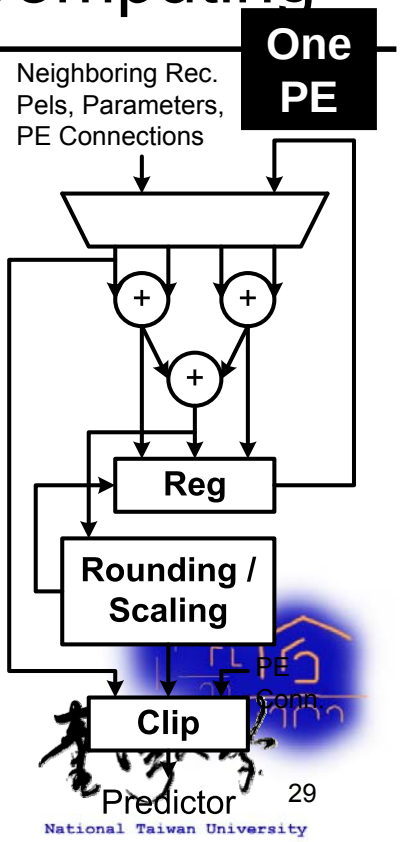
16x16 Intra Prediction (I16MB)



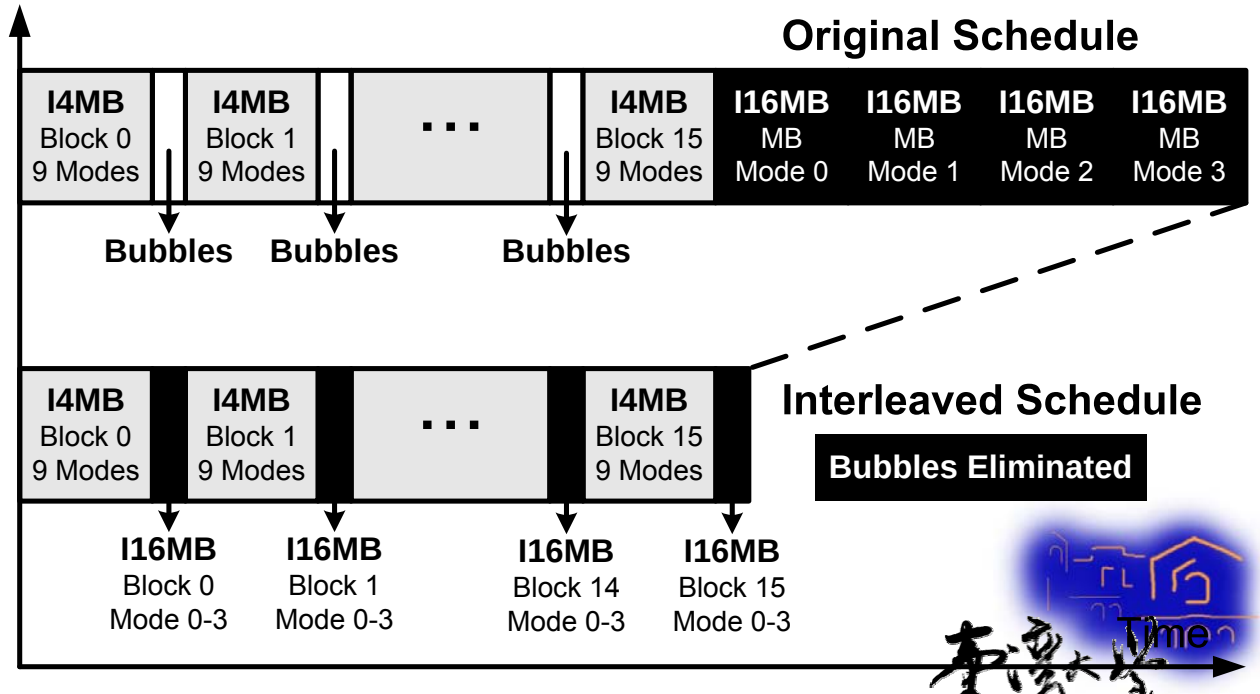
National Taiwan University

Reconfigurable Computing

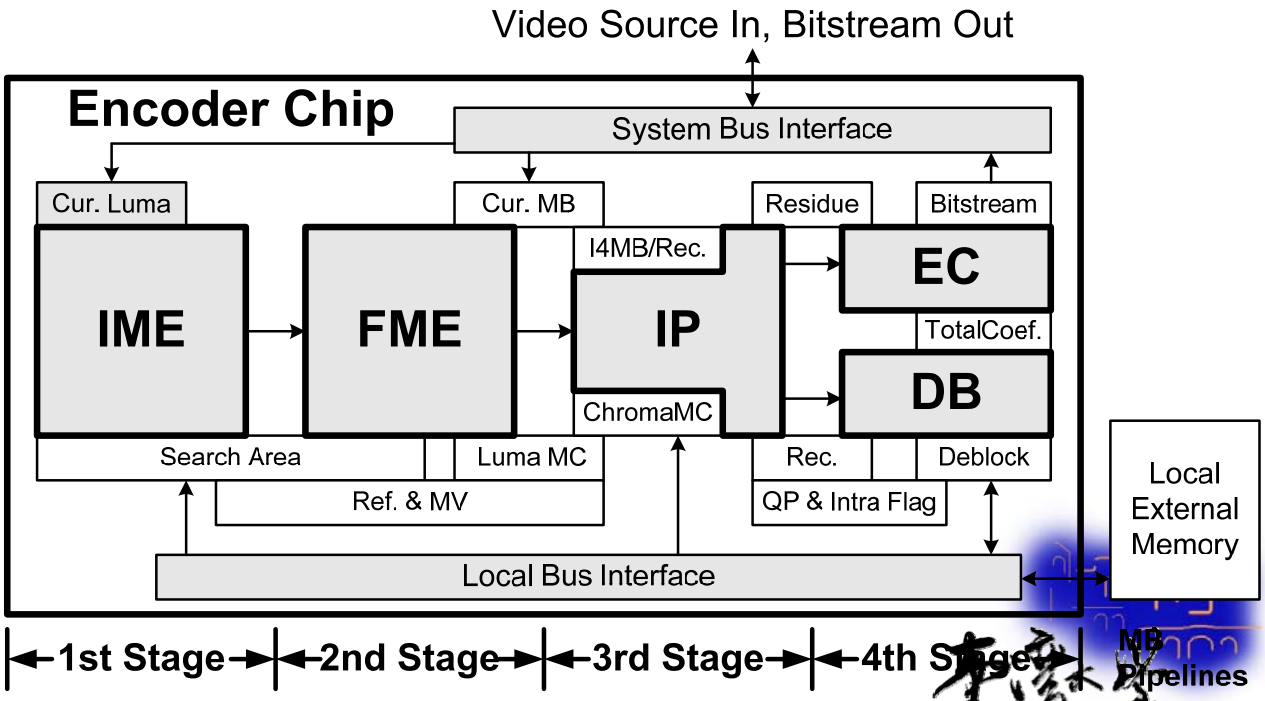
- **Four processing elements**
  - Generate four predictors in one cycle
- **Normal configuration**
  - I4MB modes except hori. and vert.
- **Bypass configuration**
  - I4MB/I16MB hori. and vert. modes
- **Cascading configuration**
  - I4MB and I16MB DC modes
- **Recursive configuration**
  - I16MB plane prediction mode



Interleaved I4MB/I16MB Schedule

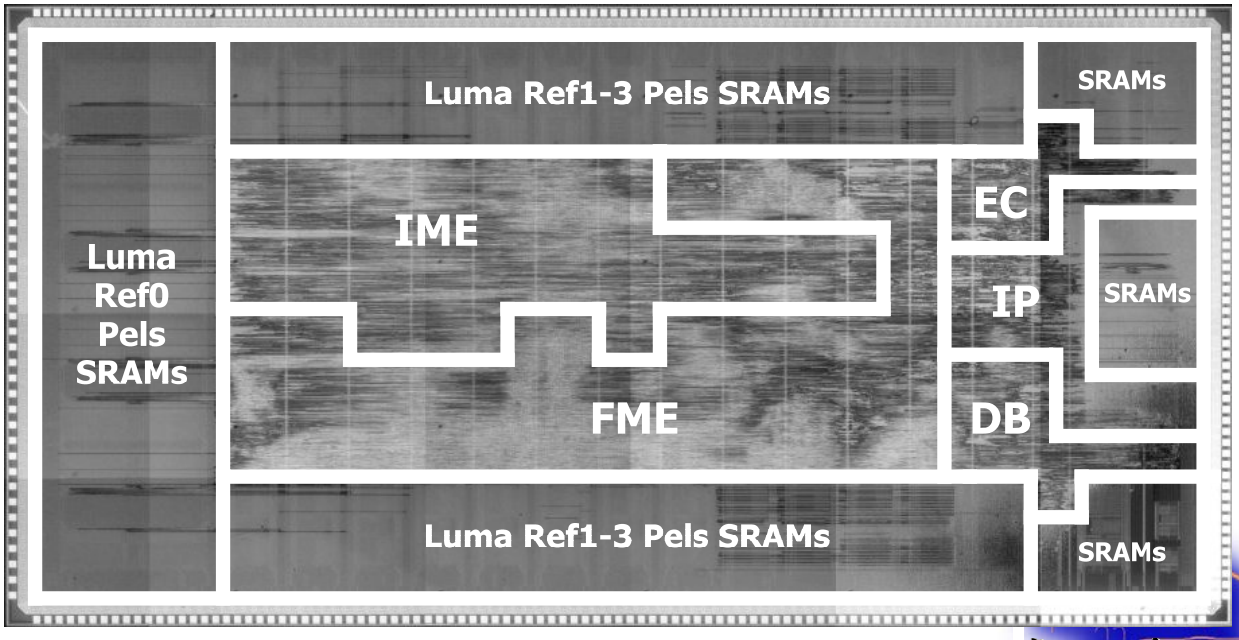


System Architecture for Video Encoding



National Taiwan University

Chip Micrograph

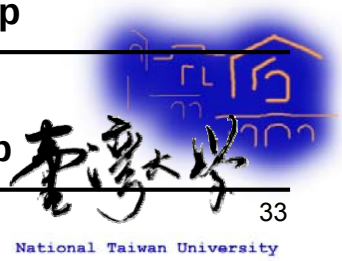


Source: Huang, NTU, ISSCC 2005

National Taiwan University

Chip Features

Technology	**0.18 μm CMOS 1P6M**
Supply Voltage	**1.8V**
Core Area	**7.68×4.13mm²**
Logic Gates	**922.8K (2-input NAND gate)**
SRAM	**34.72KB**
Encoding Tools	**Baseline Profile Compression**
Operating Frequency	**81MHz for D1** **108MHz for HDTV720p**
Power Consumption	**581mW for D1** **785mW for HDTV720p**



National Taiwan University

Outline

- Introduction
- Low power stream video processor
  - Data parallel optimazation
- Multi-core stream processor SoC for Graphic and computer vision
  - Streaming data model
- Conclusion

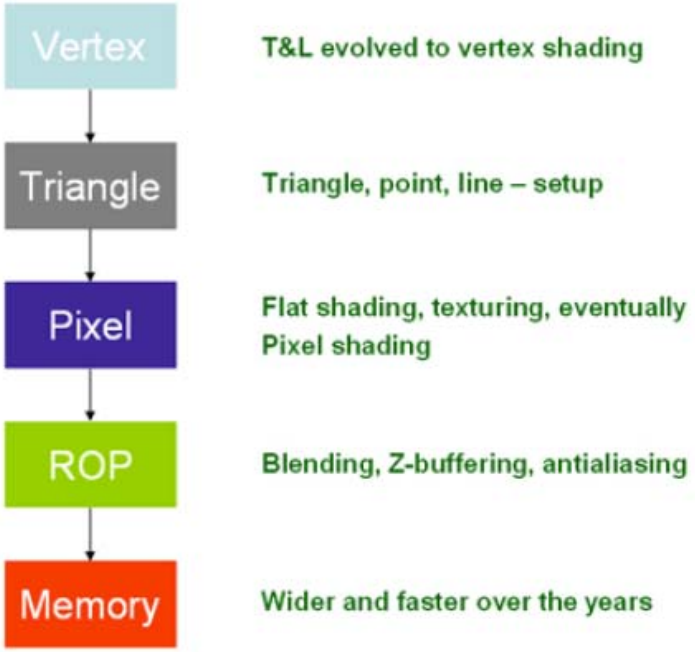
National Taiwan University

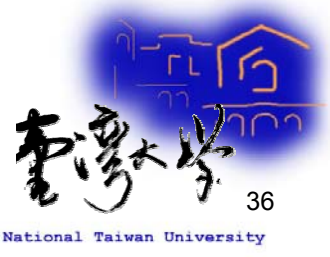
GPUs evolution

- Evolution of the PC hardware graphics pipeline:
  - 1995-1998: Texture mapping and z-buffer
  - 1998: Multitexturing
  - 1999-2000: Transform and lighting
  - 2001: Programmable vertex shader
  - 2002-2003: Programmable pixel shader
  - 2004-2006: Shader model 3.0 and 64-bit color support
  - 2007: Stream computing

National Taiwan University

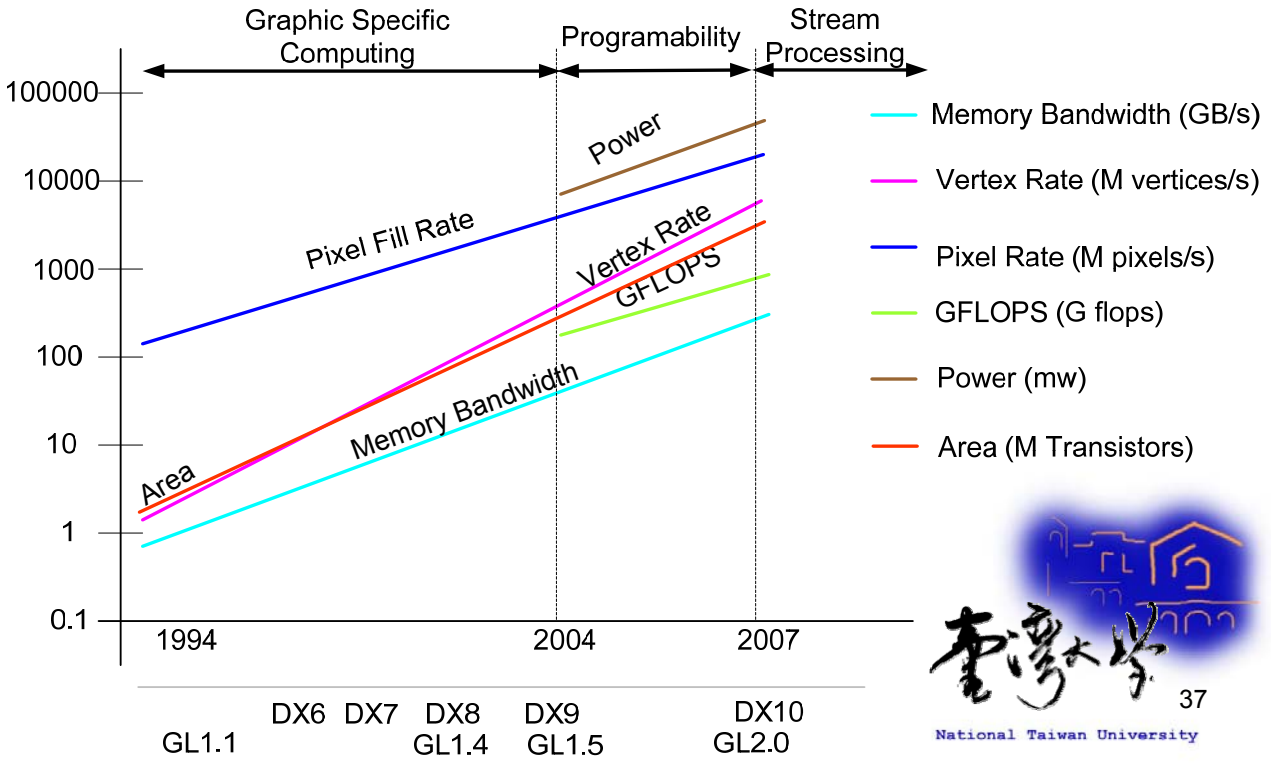
Traditional GPU pipeline





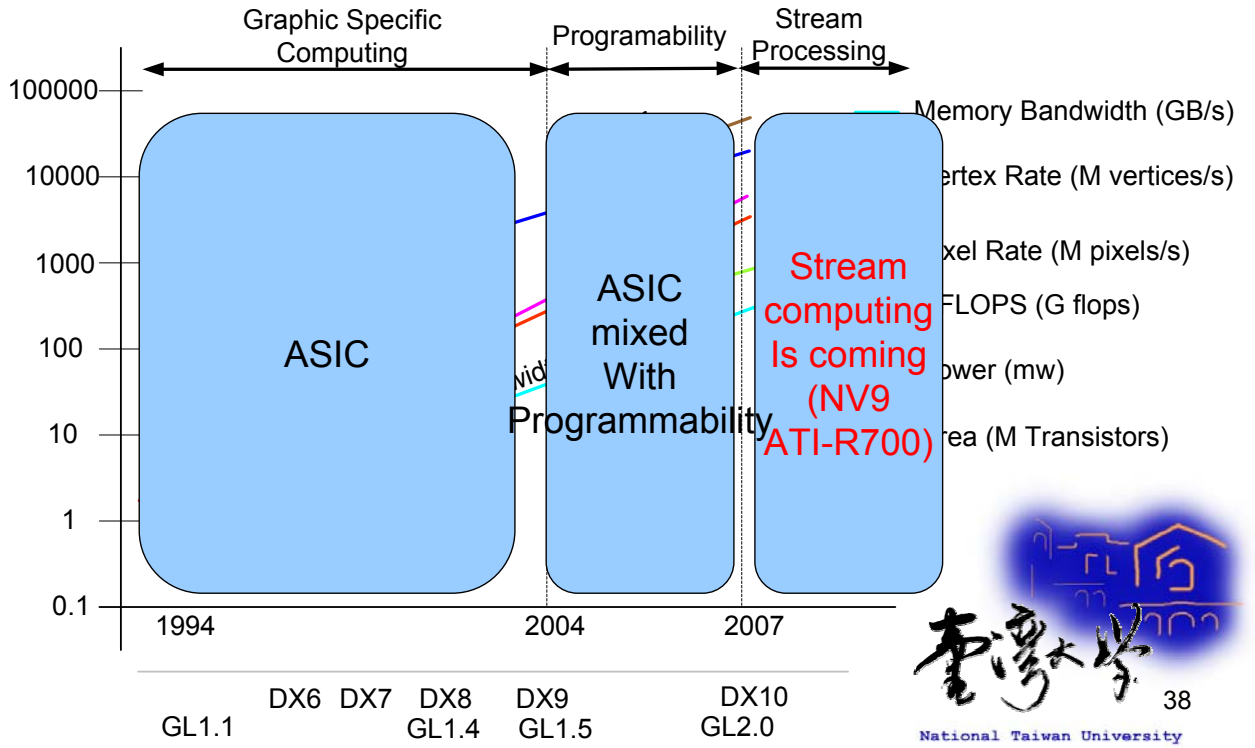
National Taiwan University

Performance evolution



National Taiwan University

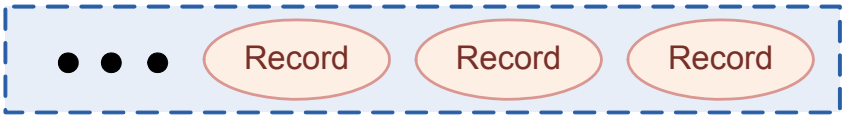
Contemporary GPUs evolution



National Taiwan University

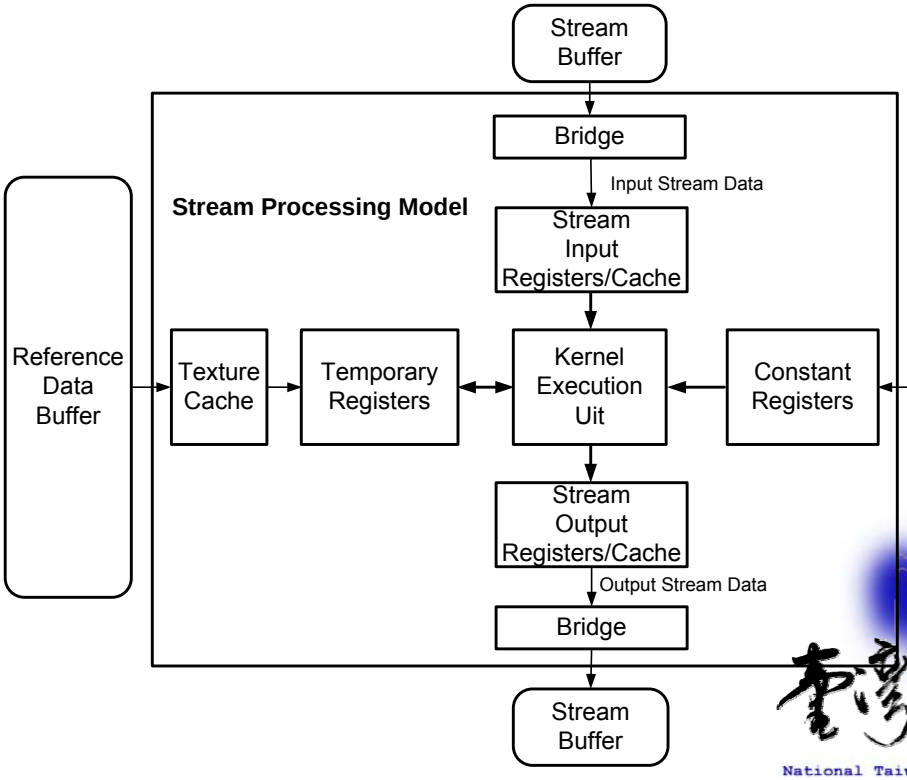
Stream Programming Model

- Stream program organizes data as streams and computation as kernel
  - Stream element (record)
    - 8-bit pixel
    - User defined data structure (MB, RGB_pixel…)

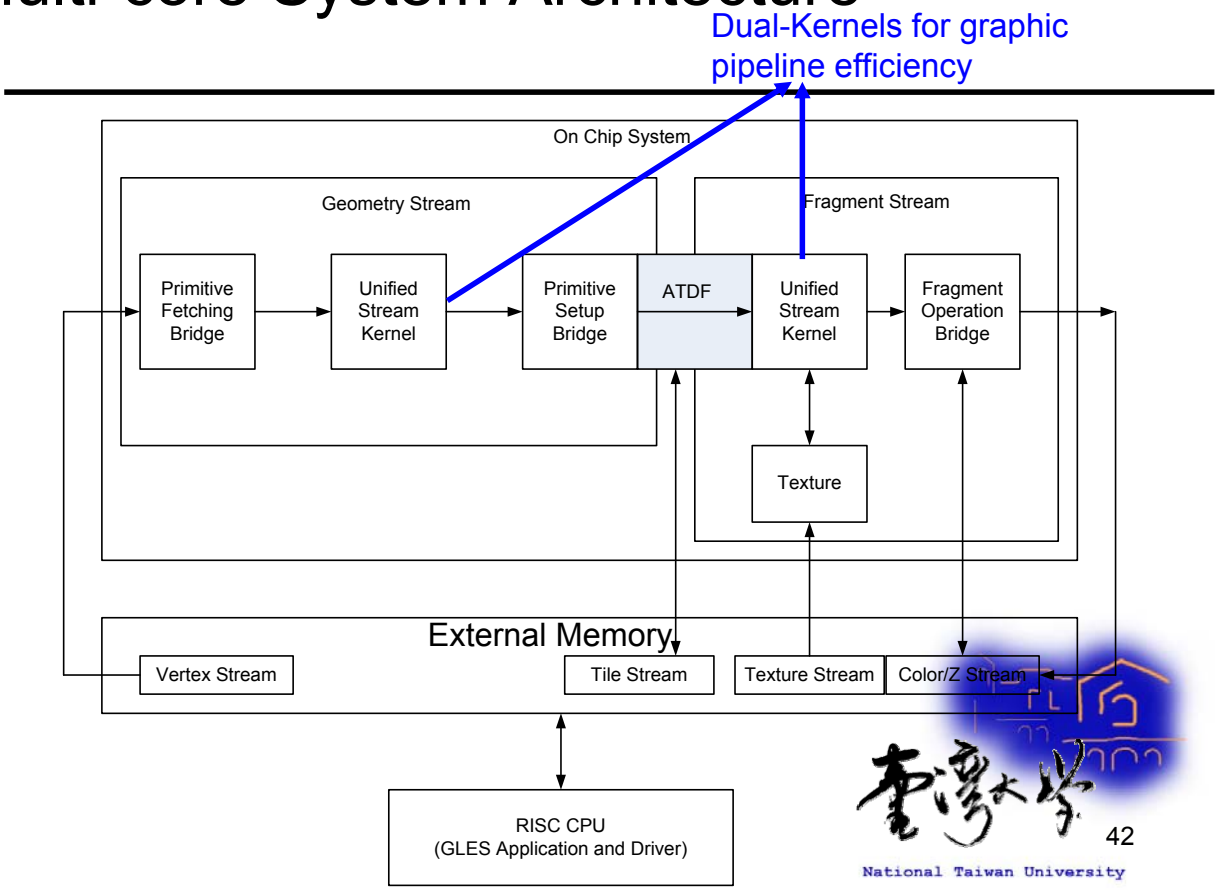


  - Kernel
    - Define computation from input streams to output streams
    - Only access local memory space!!

Stream Processing Model

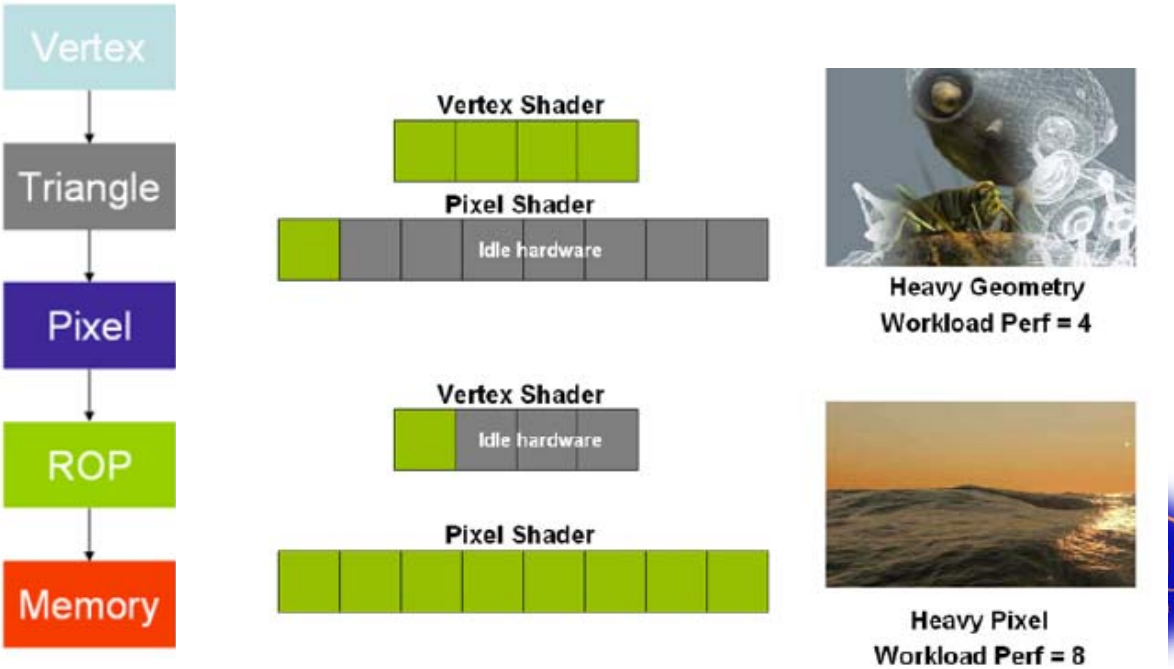


Multi-core System Architecture



National Taiwan University

Pipeline Task Sharing by Reconfigurable Stream Core



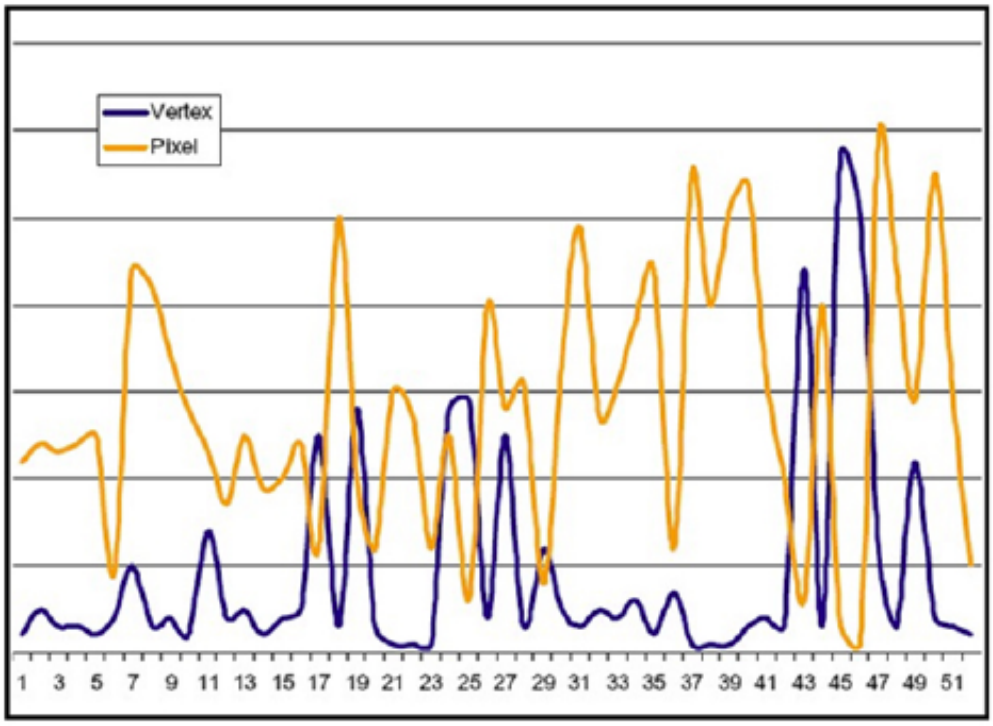
National Taiwan University

Unified Stream Core

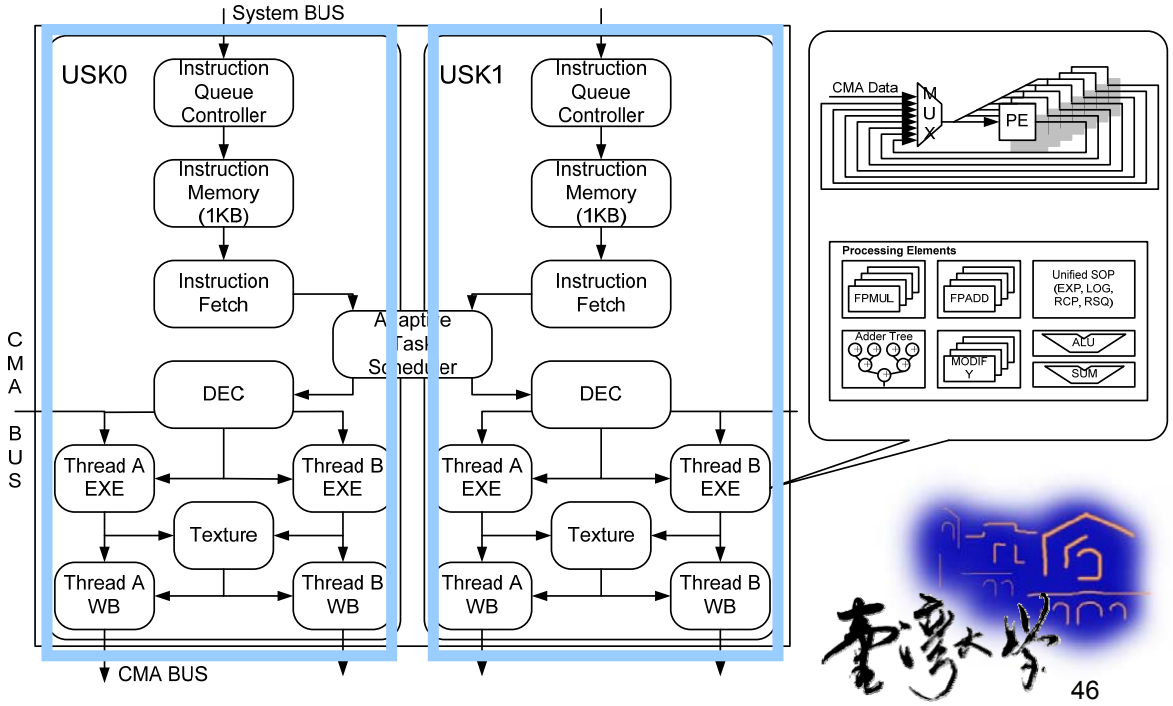
Heavy Geometry
Workload Perf =12

Heavy Pixel
Workload Perf = 12

Work Load Analysis

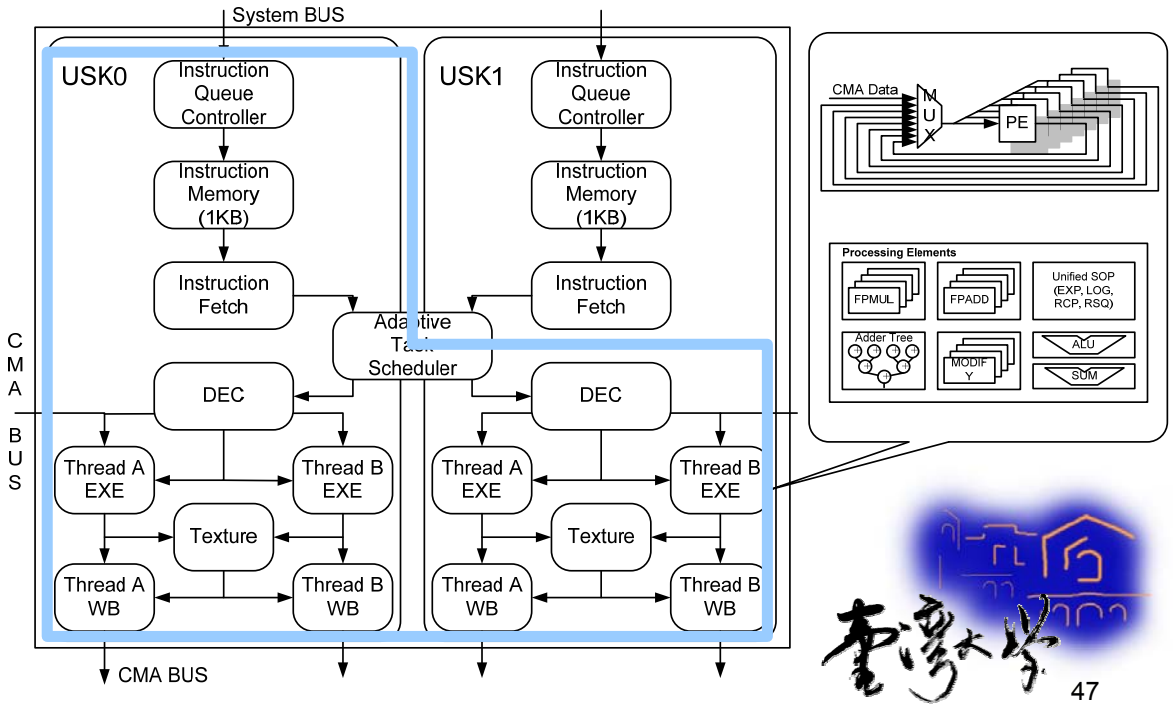


Task level scalable for multi-core system



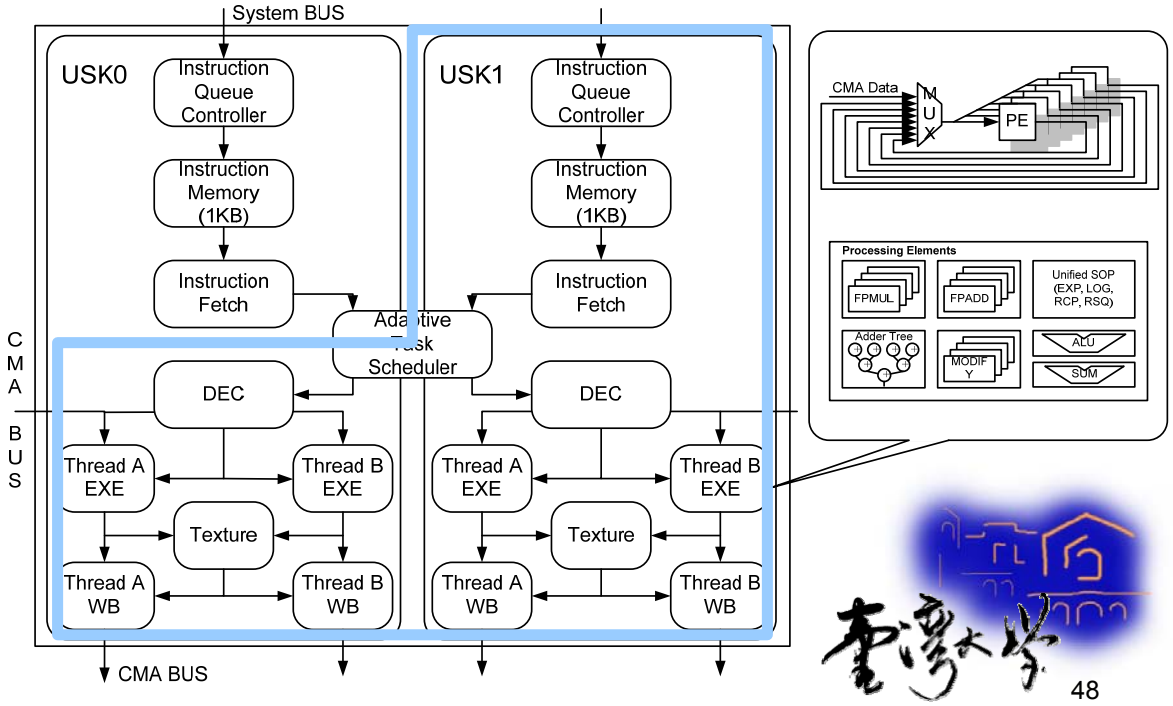
National Taiwan University

Reconfiguration for instruction path



National Taiwan University

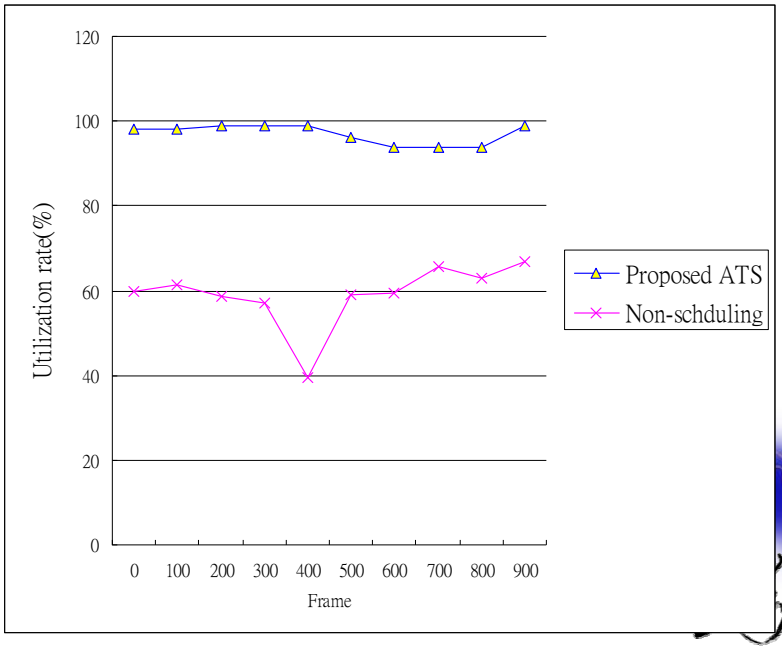
Reconfiguration for instruction path



National Taiwan University

Measurement result

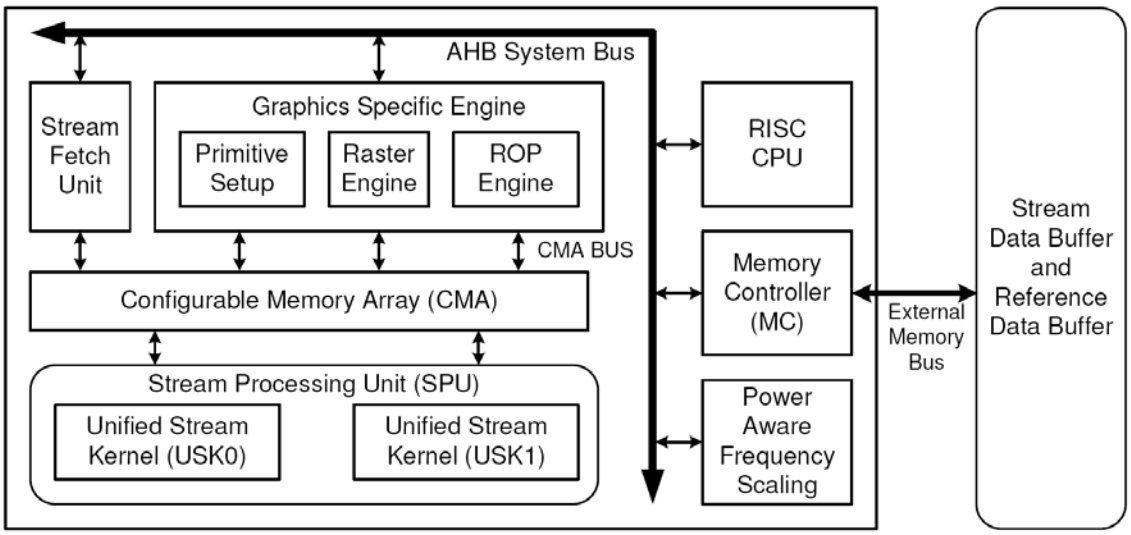
- Average 1.6 times speed up



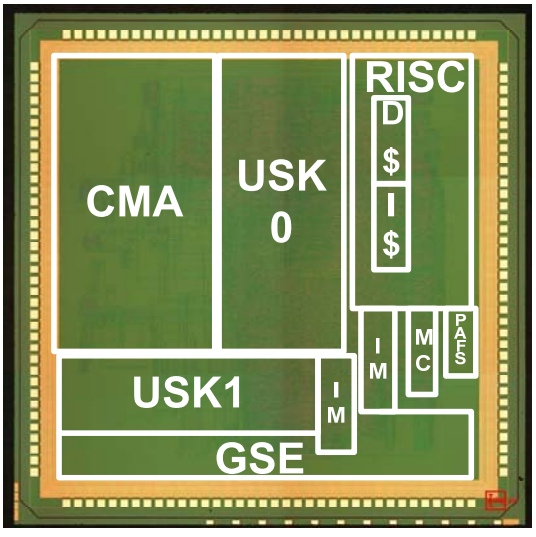
Accepted as "Configurable Memory Array for Multimedia Multithreading Stream Processing Architecture ," in Proc. IEEE International Symposium on Circuits and Systems (ISCAS'08), May 2008

Stream architecture for Mobile Multimedia

- Duo heterogeneous cores
- Due homogenous stream processor cores
- Multi-Clock domain power optimization



Die photo and specification

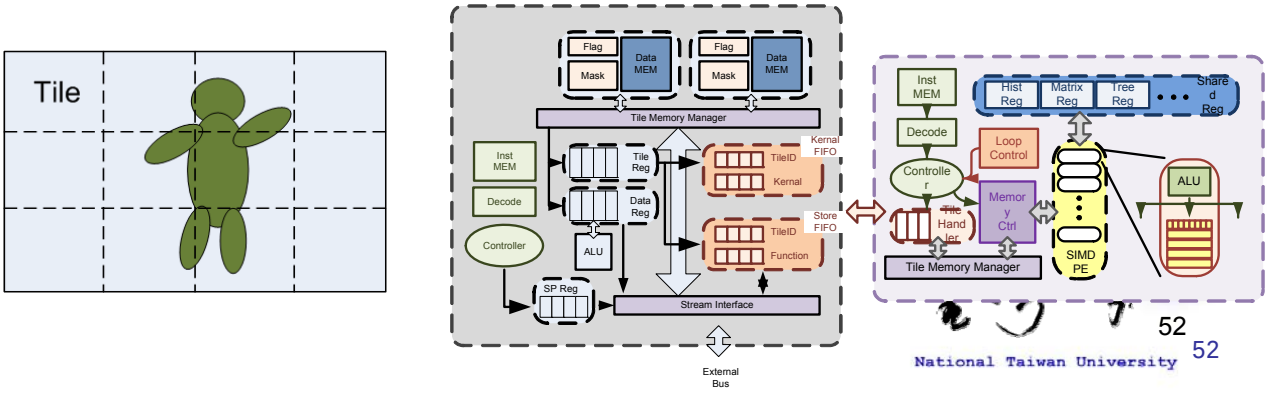


Process Technology		UMC 90nm CMOS 1P9M LowK
Supply Voltage		1.0V core, 2.5V I/O
Clock Frequency		50-200MHz, 5 CLK Domains
Power Consumption		26mW (Stream Processing Unit 16 mW)
SRAM	RISC CPU	I$: 8KB D$: 8KB
	SPU	IM: 2KB
	CMA	10KB
Performance	Arithmetic	16 GOPS
		6.4 GFLOPS
	Graphics Throughput	200M vertics/s, 400M pixels/s

Accepted as "A 26mW 6.4GFLOPS Multi-Core Stream Processor for Mobile Multimedia Applications," in Digest of Technical Papers of Symposium on VLSI Circuits, 2008.

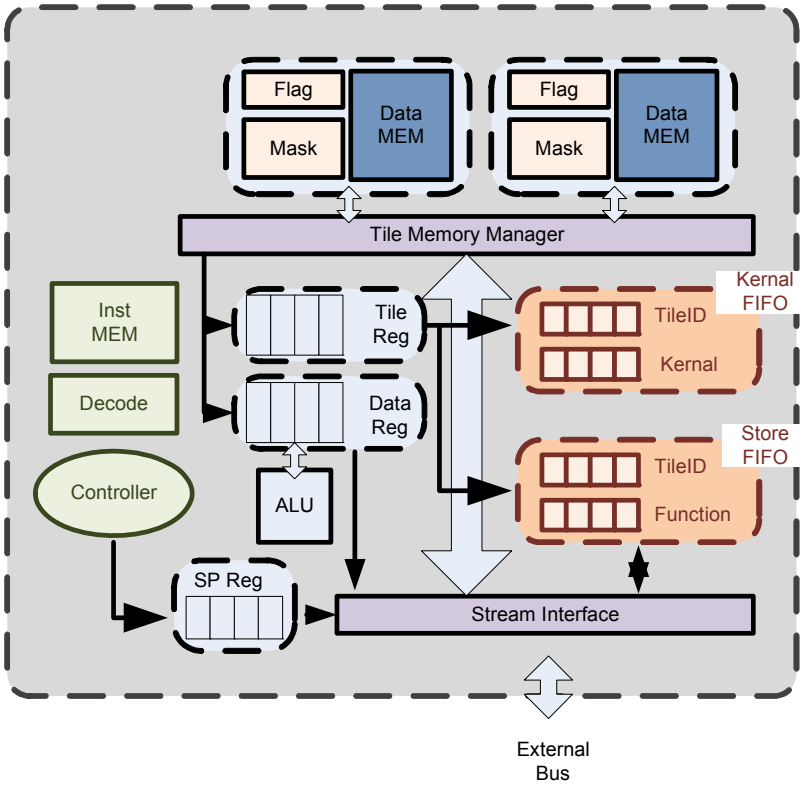
Stream Architecture for Computer Vision

- 2D Stream processor
  - Computer vision function architecture
  - Tile base stream processing
- Parallel mask conditional operation
  - Avoid a lot of branch instructions



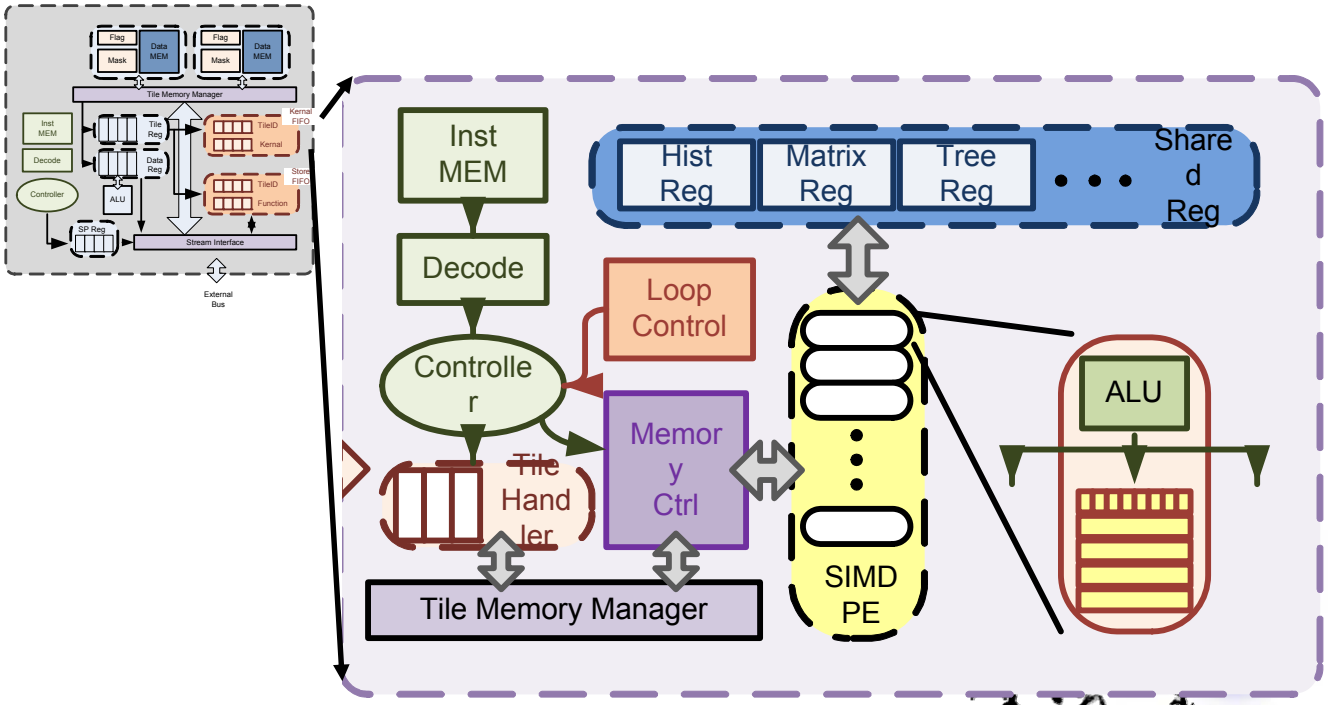
52

Stream Architecture for Computer Vision

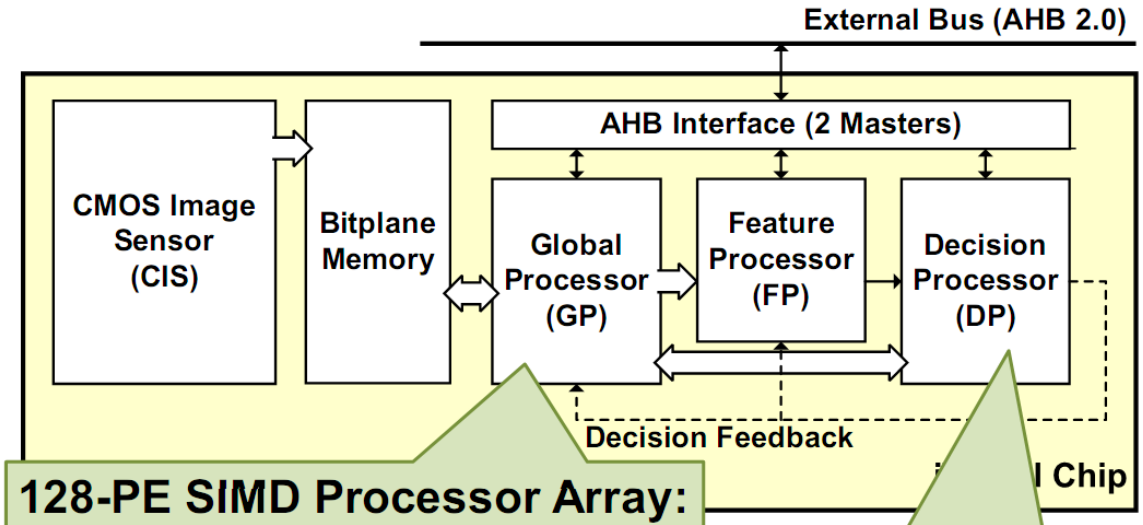


53

Stream Architecture for Computer Vision

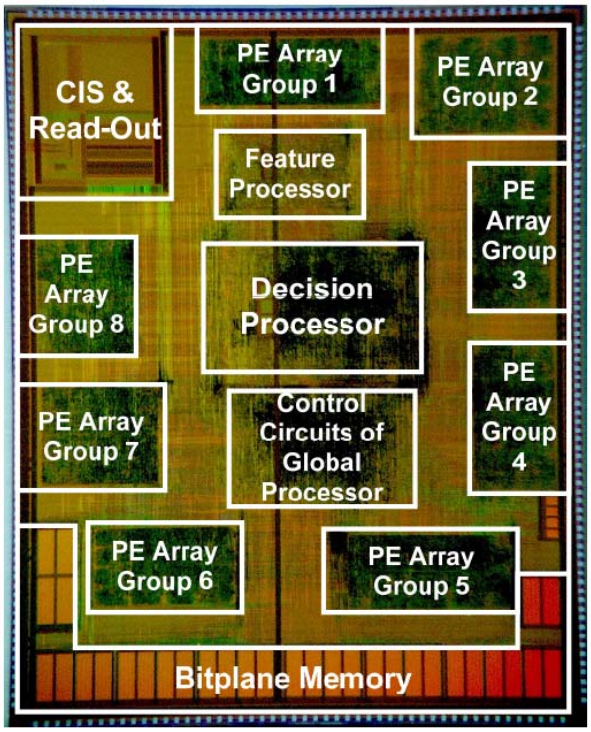


Scalable Visual Processor



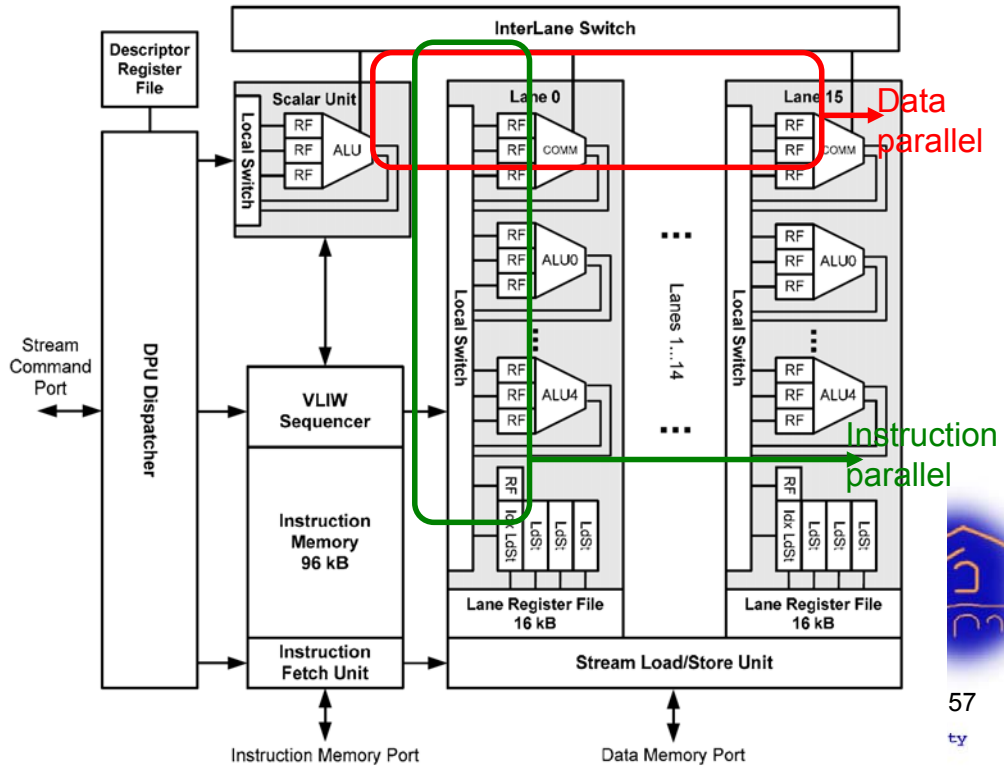
iVisual (Intelligent Visual Processor)

- Highest level abstraction
  - Image-in, answer-out
- Scalable architecture for higher specification
  - Can be easily combined
  - Change integrated PE array number



Source: Chin, NTU, ISSCC 2008

Video Stream Processor

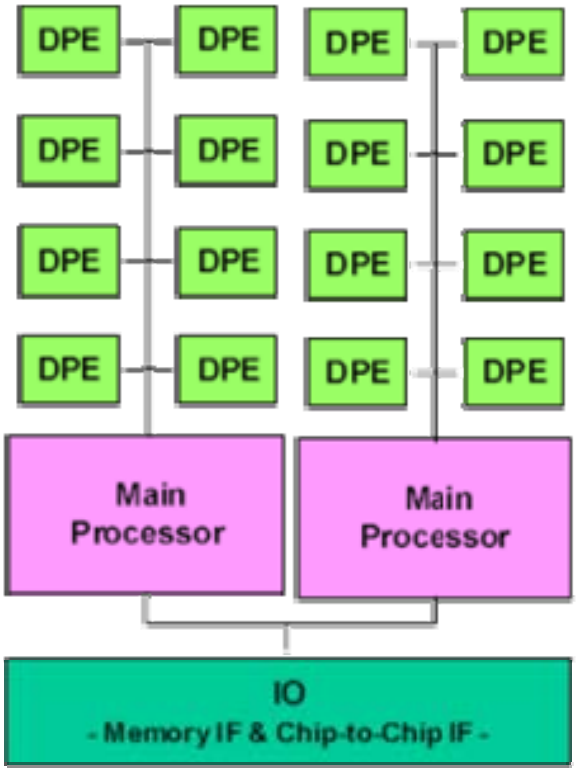


Source: ISSCC 2007

Architectural Perspective for Multimedia Processing

- Parallel Processing
  - Requirement: Processing of independent segments
  - Pipelining, systolic processing
  - Task level parallelism
  - Instruction level parallelism
  - Data level parallelism
- Stream Processor
  - Requirement: Processing of dependent segments
  - Scalable architecture
  - Multi-thread and cache

Architectural Perspectives of MPSOC



Stream base PE Core with:
−Data parallel
−Instruction parallel
−Reconfigurable
−Scalable

Generic CPU with
-Record preparing

Conclusion

- Semiconductor Technology and Mobile MultiMedia applications are continuously the push-pull driving forces for MPSoC.
- End Products with Battery power for MMM requires novel multi-core architectures.
- Stream Processor could play important role to provide reconfigurability and scalability.

Thank you!