Computational Platforms for Virtual Immersion Architectures:

Integrating Specialized Processing Units Into A Homogeneous Array of Processors

Bevan Baas

Department of Electrical and Computer Engineering
University of California, Davis

July 10, 2008
SRC/NSF/ITRS 2020 VIA Forum

Students

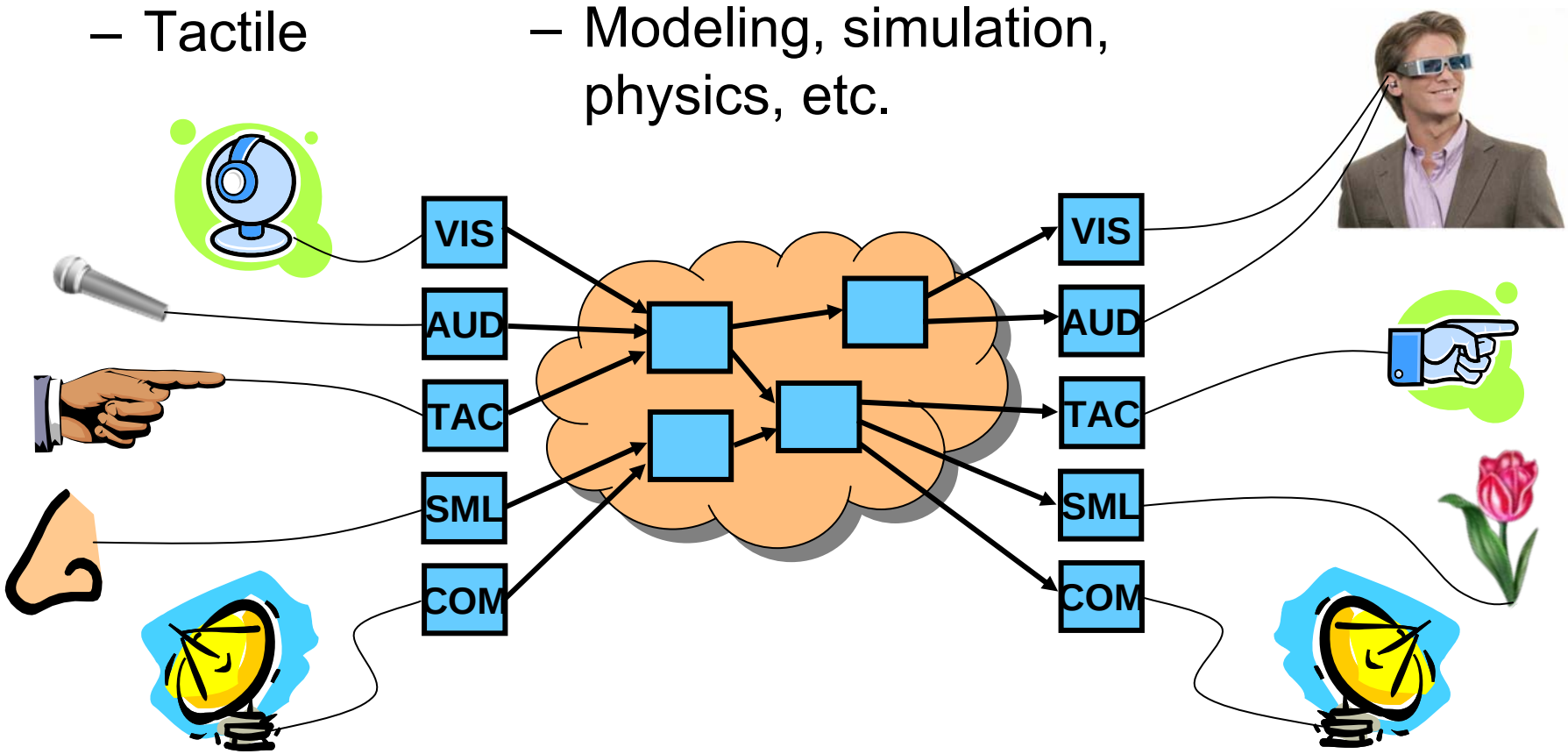
- Graduate
  - Tinoosh Mohsenin
  - Dean Truong
  - Zhibin Xiao
  - Paul Mejia
  - Anh Tran
  - Jeremy Webb
  - Christine Watnik
  - Gouri Landge
  - Jeff Collins
  - Stephen Le
- Undergraduate
  - Jon Pimentel
  - Gary Chung
  - Henna Huang
  - Kyle Piper
- Alumni
  - Zhiyi Yu — Ph.D. ’07
  - Wayne Cheng — MS ‘08
  - Eric Work — MS ‘07
  - Anthony Jacobson — MS ‘07
  - Michael Meeuwsen — MS ‘05
  - Ryan Apperson — MS ‘04
  - Michael Lai — MS ‘04
  - Omar Sattari — MS ‘04

Outline

- Virtual Immersion Applications and Hardware
- AsAP project
  - Goals
  - Hardware
    - Special-Purpose Processors
    - Inter-processor Interconnect
  - Applications

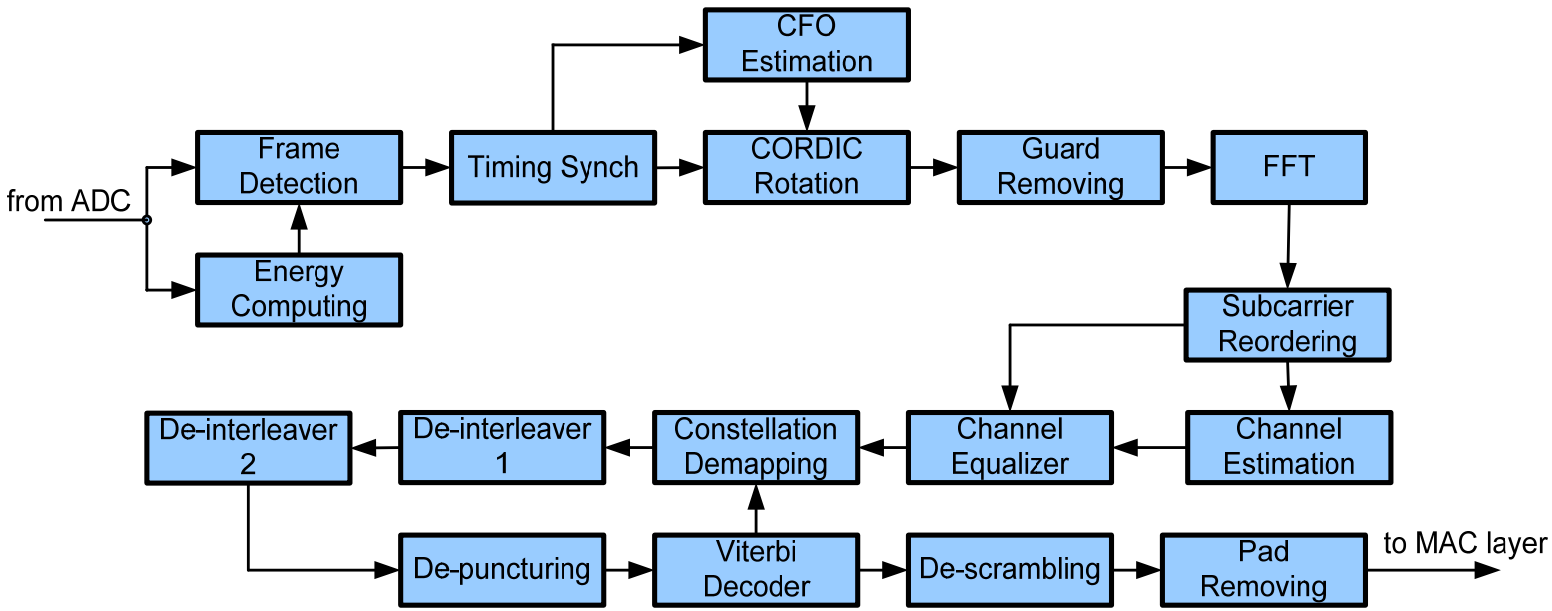
2020 Virtual Immersion

- Multiple simultaneous complex applications
  - Visual
  - Audio
  - Tactile
  - Smell
  - Communications
  - Modeling, simulation, physics, etc.



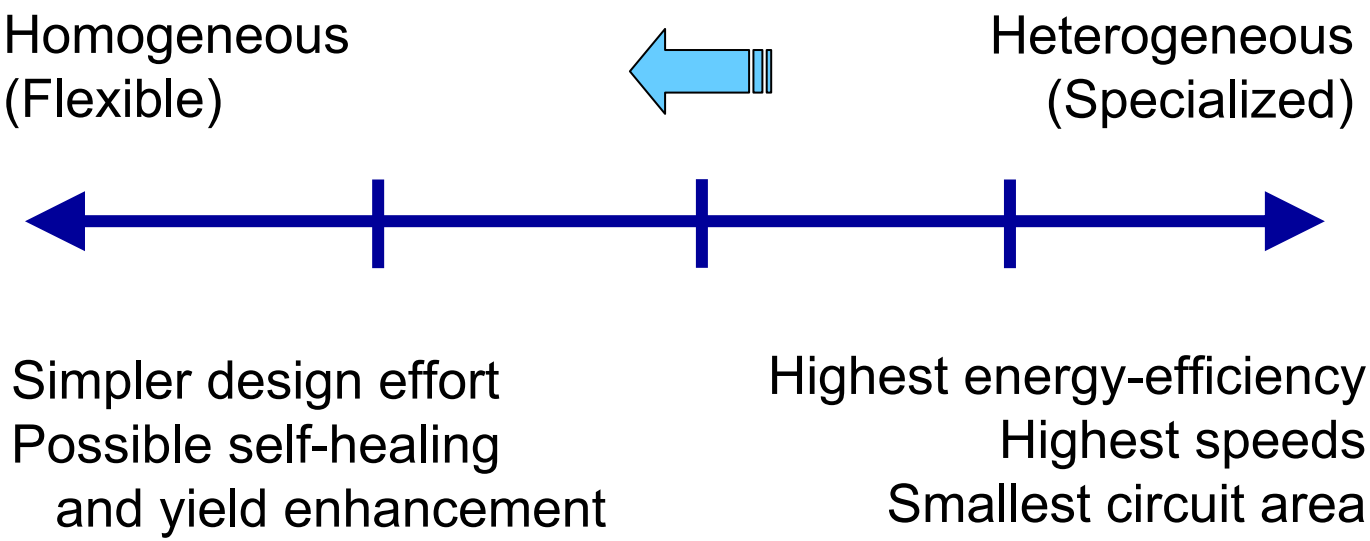
2020 Virtual Immersion

- Many of these applications are not typical general purpose
  - Involve Digital Signal Processing
  - Computationally intensive
  - Real-time
  - Composed of multiple “simple” tasks
  - Example: IEEE 802.11a/g WiFi wireless LAN receiver



Architectural Goals

- Targeting digital signal processing and multimedia workloads
- Homogeneous vs. Heterogeneous



Hardware Goals

- Programmable and reconfigurable
  - Well matched to a “broad range” of multi-task applications
- High energy efficiency (active or inactive)
- Capable of high performance
- Well suited for future fabrication technologies
  - Billions of transistors
  - Large variations across chip
  - Usable for a wide range of applications (mitigate high NRE costs)
  - Tile-able architecture (reduces design costs)
  - Many faulty circuits across a chip

Asynchronous Array of Simple Processors (AsAP)

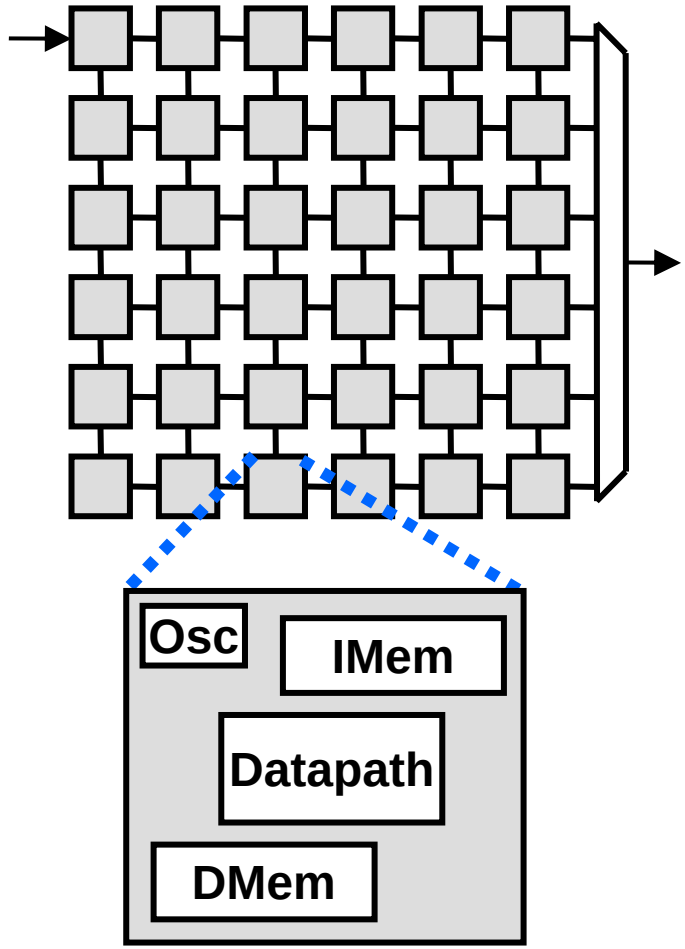
- Key Ideas:
  - Programmable, small, and simple fine-grained homogeneous cores with limited accelerators
  - Small local memories sufficient for simple kernels

"You know you have achieved perfection in design,
not when you have nothing more to add,
but when you have nothing more to take away."

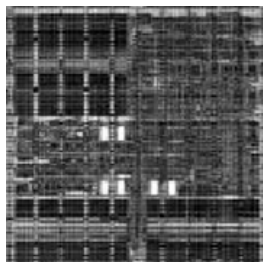
- Antoine de Saint-Exupéry

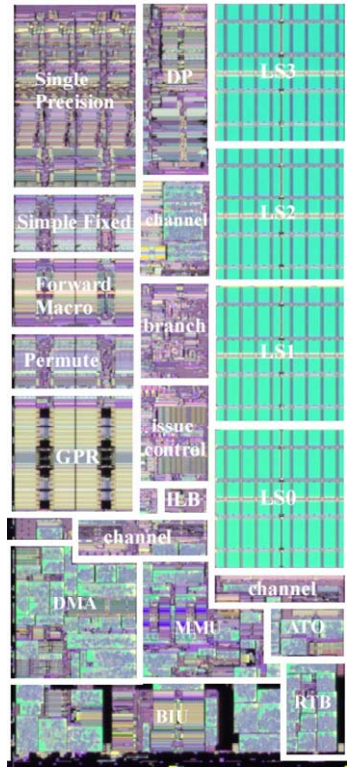
Asynchronous Array of Simple Processors (AsAP)

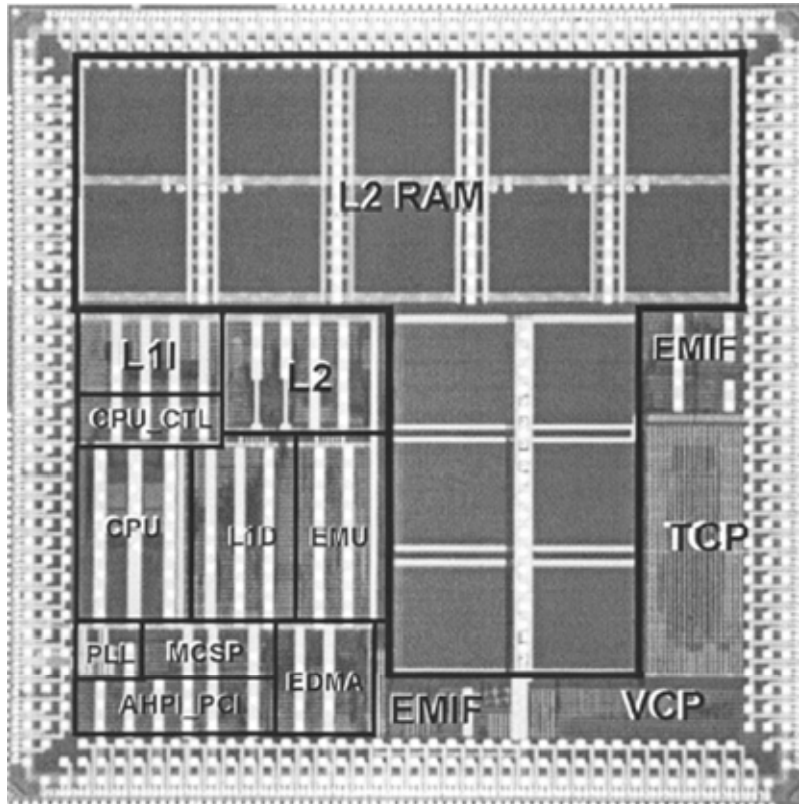
- Key Ideas, con't:
  - 2D mesh, circuit-switched inter-processor communication
    - Nearest-neighbor comm. "only"
    - Low area overhead
    - Easily scalable array
  - Globally Asynchronous and Locally Synchronous (GALS) clocking
    - Independent oscillator and clock frequencies on every core
    - Local oscillator immediately halts when processor is idle



Area Comparison





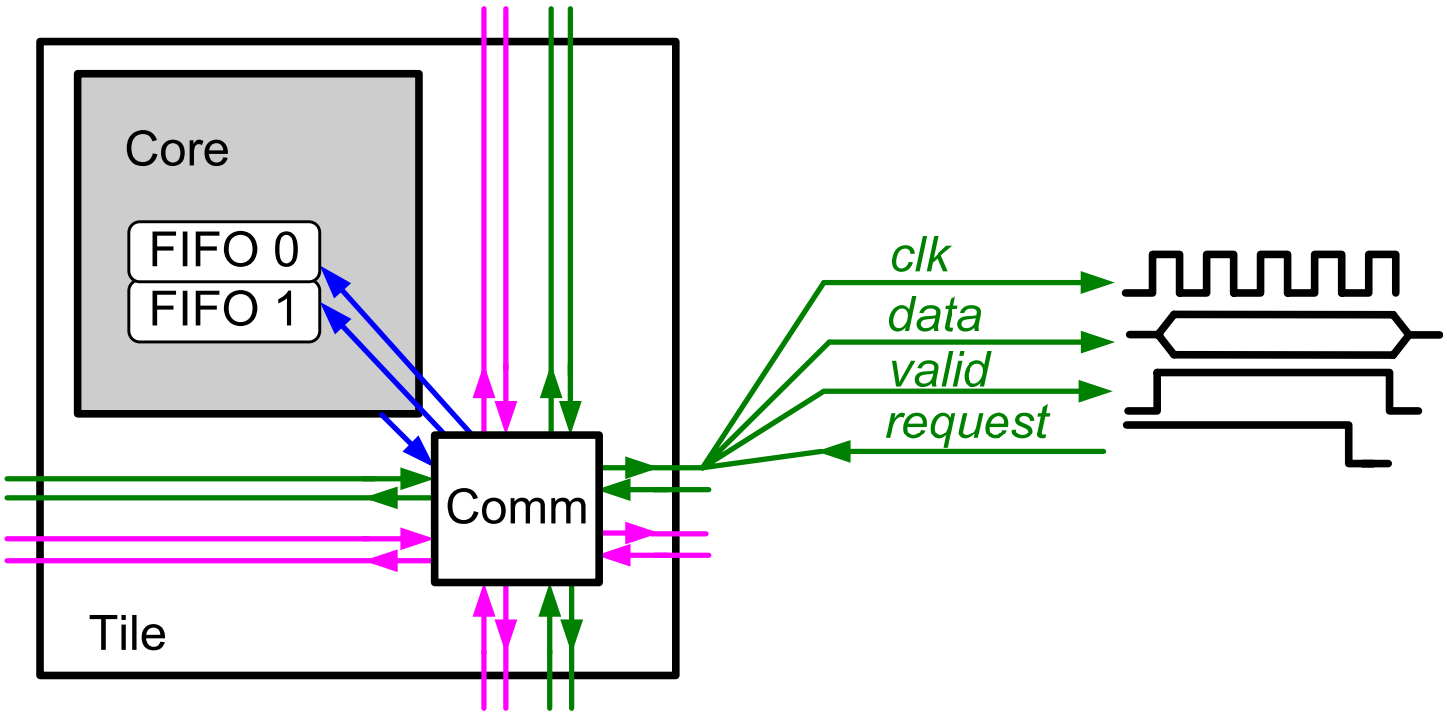


AsAP **RAW** **CELL SPE** **TI C64x**

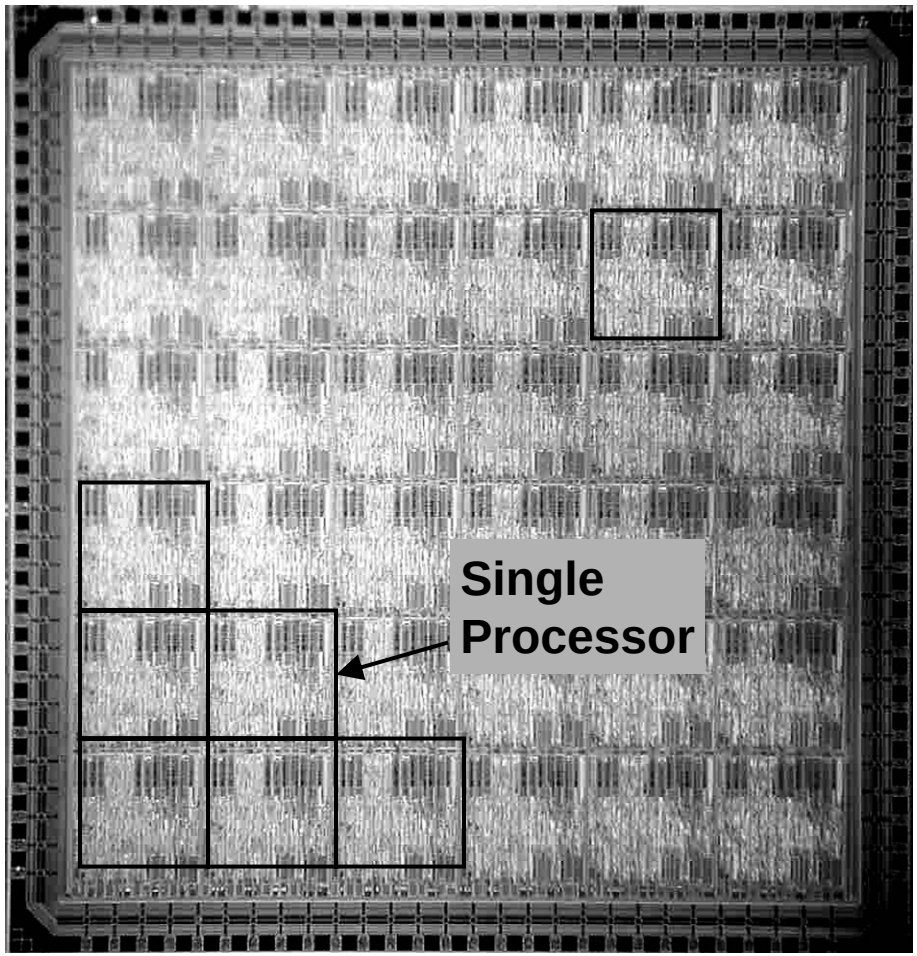
- Approximate scale die micrographs of multi-core processors (AsAP, RAW, CELL SPE), and the TI C64x DSP processor scaled to the same technology

Inter-Processor Communication

- Circuit-switched source-synchronous communication
  - Each link has a *clk*, 16-bit *data* bus, *valid*, and *request*



36-Processor AsAP1



Technology: TSMC 0.18 µm

Max speed: **610 MHz** @ 2.0 V

Area: **0.66 mm²**

Power (1 Proc @ 1.8V, 475 MHz):

Typical application **32 mW**

Power (1 Proc @ 0.9V, 116 MHz):

Typical application **2.4 mW**

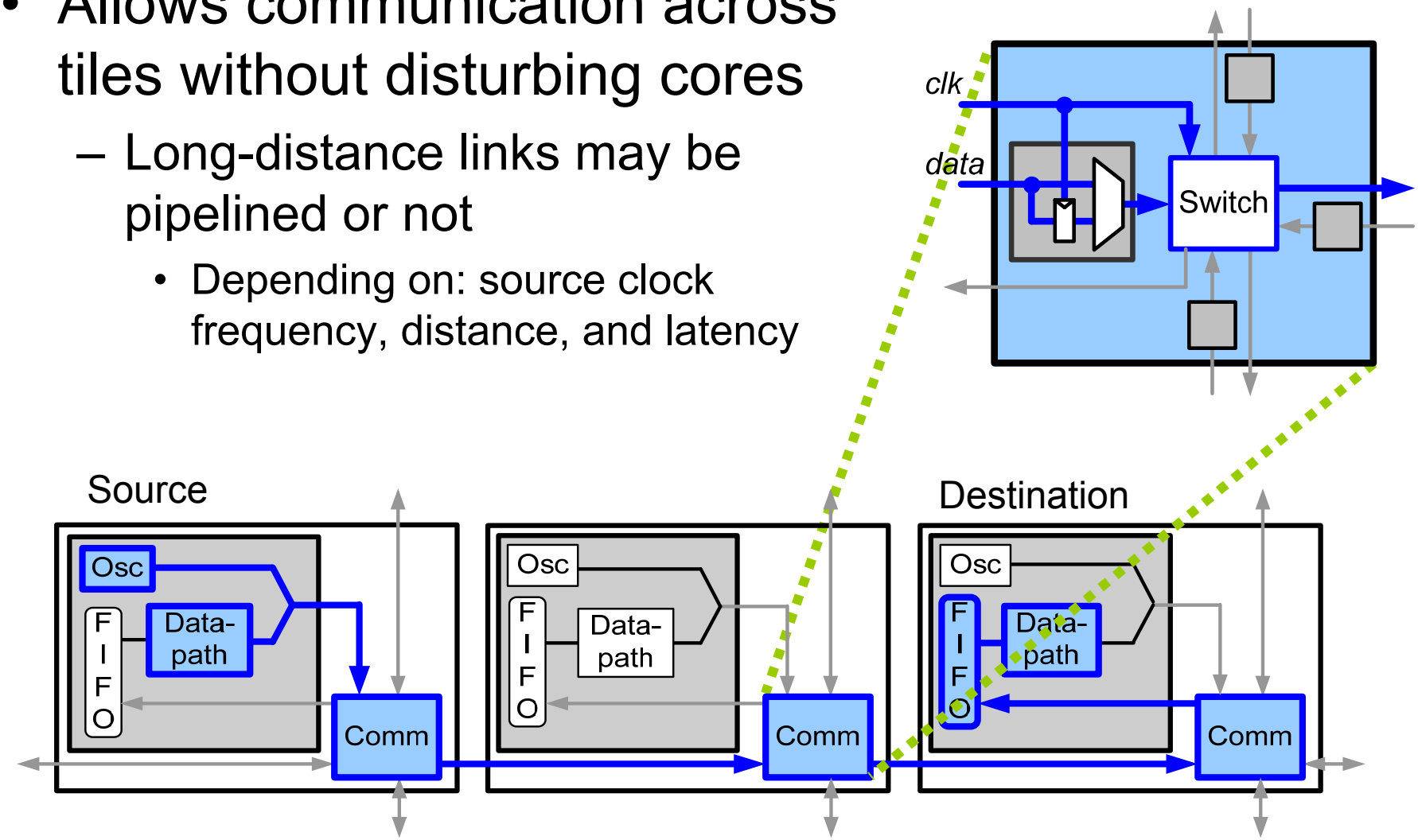
[ISSCC 06, HotChips 06, IEEE Micro 07, TVLSI 07, JSSC 08,…]

New Challenges Addressed

- Efficient, low overhead communication between distant processors
- Achieving very high efficiencies and speed on common demanding tasks such as:
  - Fast Fourier Transform (FFT)
  - Motion estimation for video encoding
  - Viterbi decoding
- Larger on-chip memories

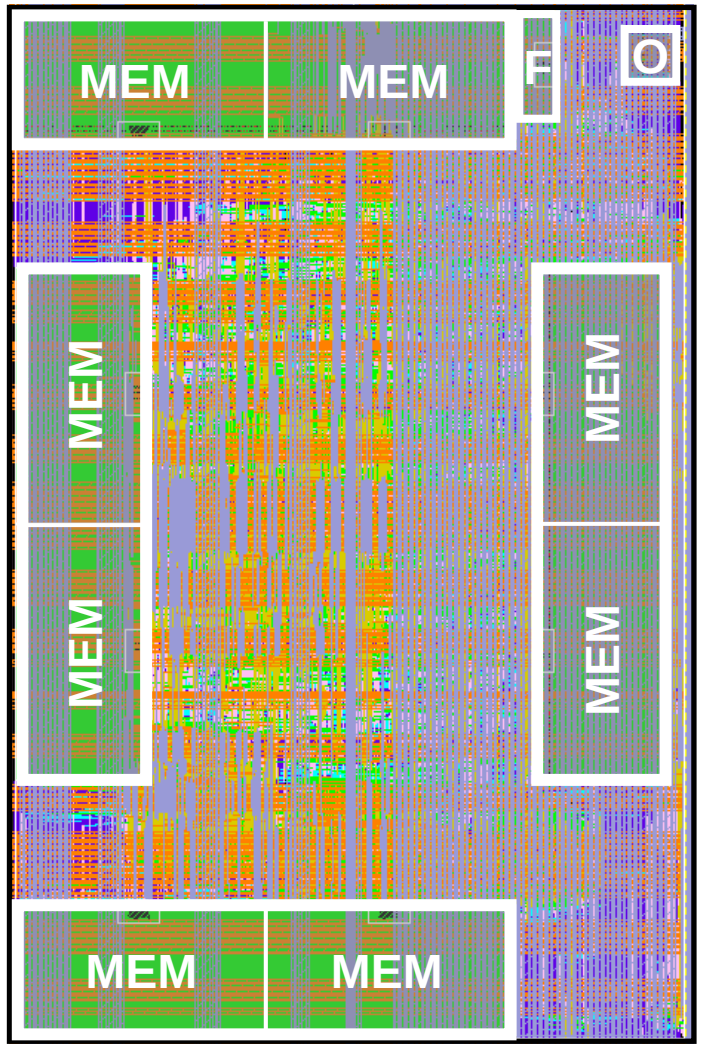
Long-Distance Communication

- Allows communication across tiles without disturbing cores
  - Long-distance links may be pipelined or not
    - Depending on: source clock frequency, distance, and latency



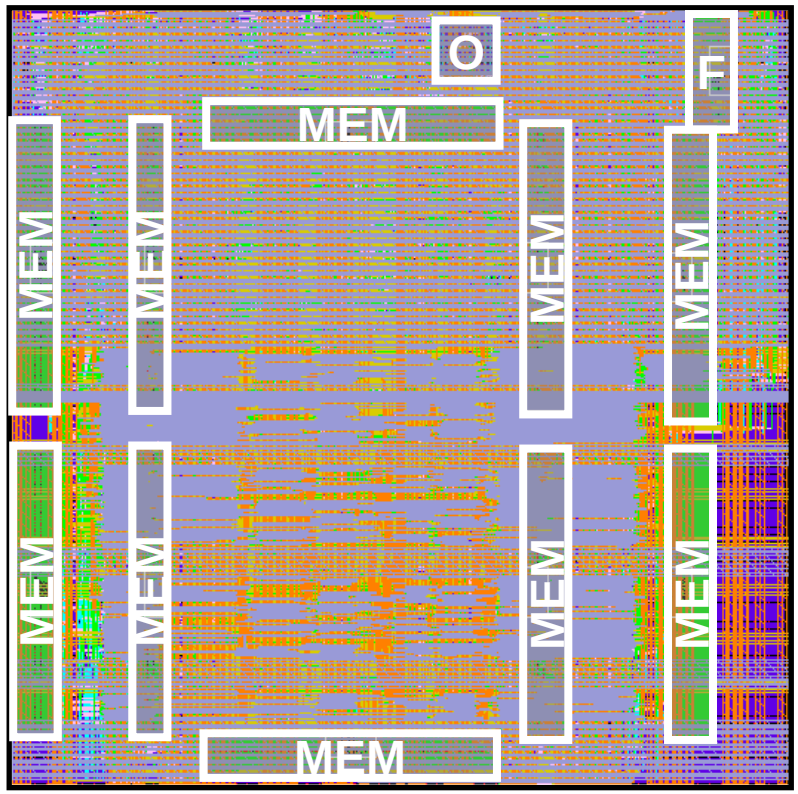
Fast Fourier Transform (FFT)

- Runtime configurable from 16-pt to 4096-pt complex transforms, FFT and IFFT
- Preliminary measurements functional at 866 MHz, 35 mW
  - 665,000 1024-pt complex FFTs/sec → 681 MSamp/s
- 1.01 mm^2



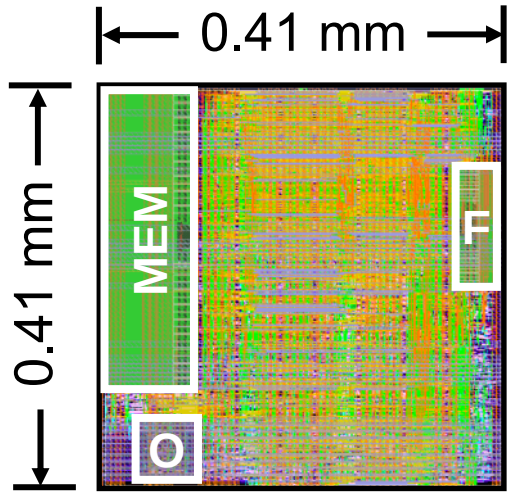
Motion Estimation for Video Encoding

- Supports all H.264 specified block sizes within a 48x48 search range
- Supports a number of fixed and programmable search patterns
- Preliminary measurements functional at 938 MHz, 196 mW
  - 15 billion SADs/sec
  - supports 1080p HDTV @ 30fps
- 0.67 mm^2



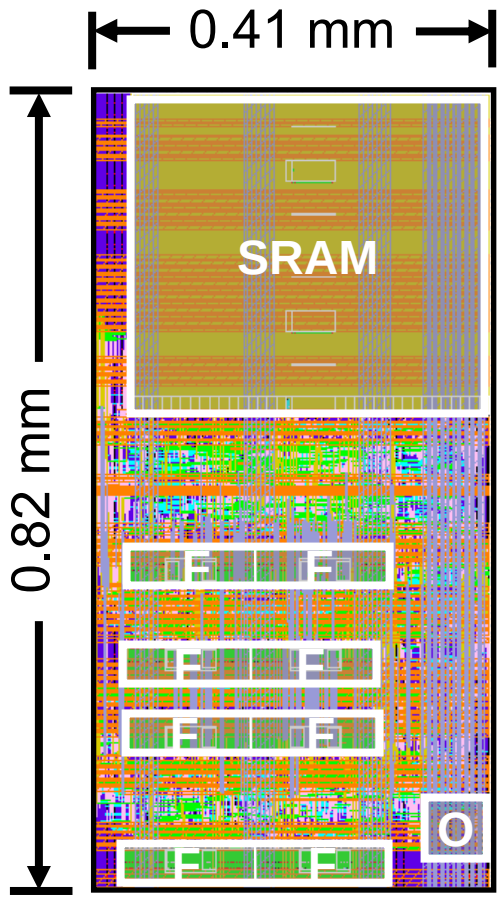
Viterbi Decoder

- 8 Add-Compare-Select (ACS) units
- Highly configurable
  - Up to 32 different rates, including 1/2 and 3/4
  - Decode codes up to constraint length 10
- Preliminary measurements functional at 894 MHz, 17.6 mW
  - 82 Mbps at rate=1/2
- 0.17 mm^2



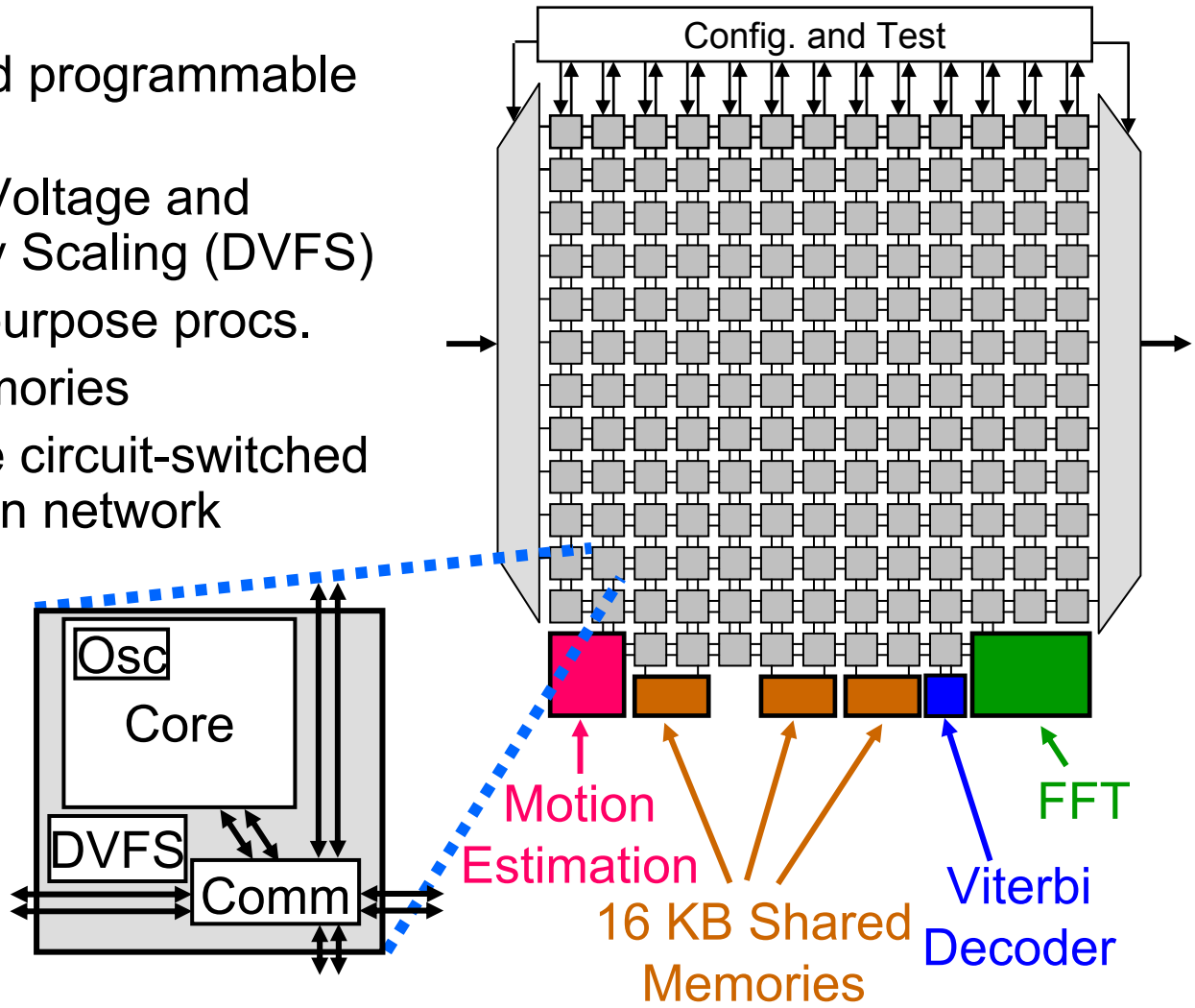
Shared Memories

- Ports for up to four processors (two connected in this chip) to directly connect to the block, which provides
  - Port priority
  - Port request arbitration
  - Programmable address generation
- 16 KByte single-ported SRAM
- Preliminary measurements functional at 1.3 GHz, 4.55 mW
  - One read or write per cycle
  - 20.5 Gbps peak throughput
- 0.34 mm^2



167-processor Computational Platform

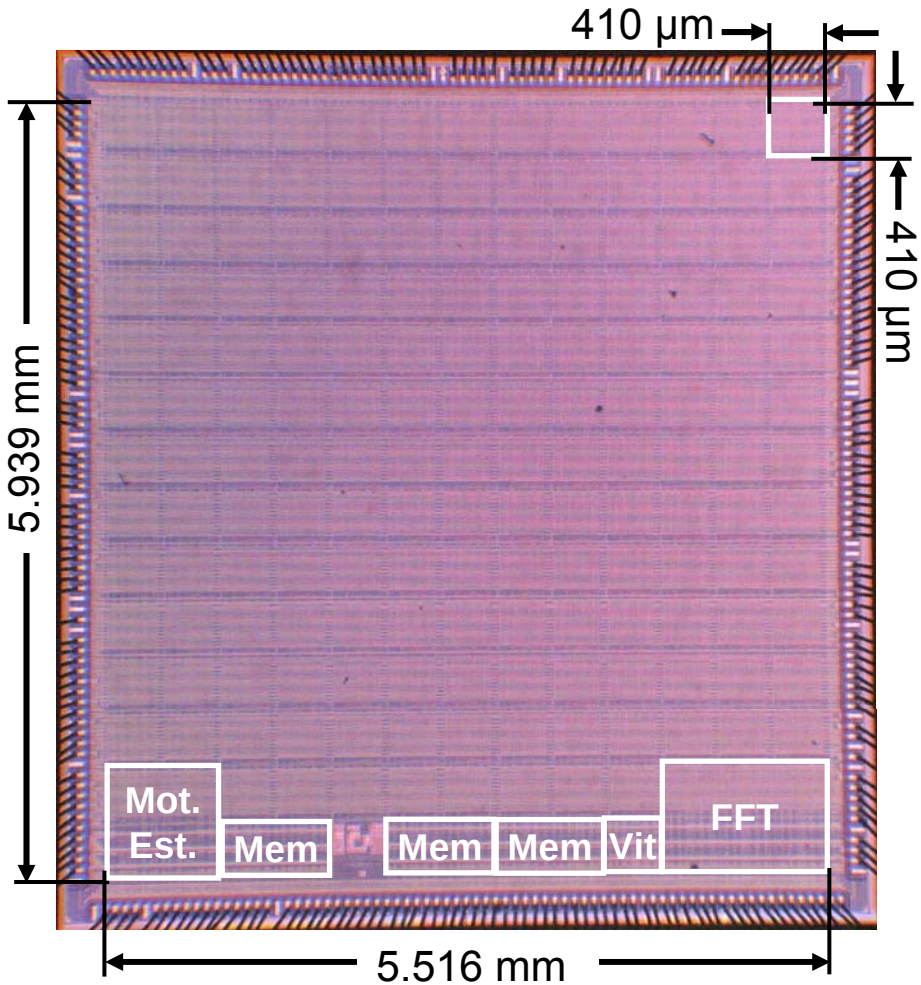
- Key features
  - 164 Enhanced programmable processors
    - Dynamic Voltage and Frequency Scaling (DVFS)
  - 3 Dedicated-purpose procs.
  - 3 Shared memories
  - Long-distance circuit-switched communication network



Die Micrograph and Key Data

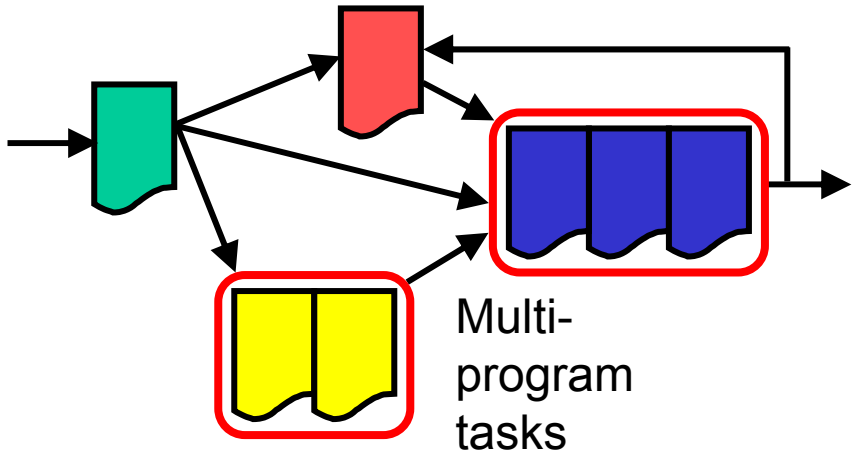
- 65 nm low-leakage CMOS
- Fully functional

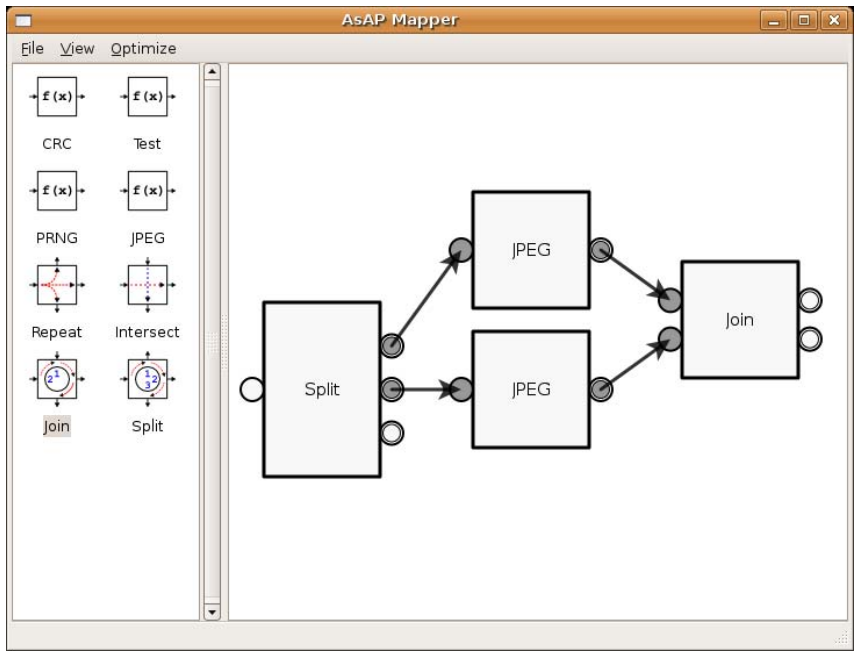
Single Tile	
Area	0.17 mm^2
Max. frequency	1.19 GHz @ 1.3 V
Power (100% active)	59 mW @ 1.19 GHz, 1.3 V
	608 µW @ 66 MHz, 0.675 V (9 pJ/op)
Power (appl.)	~20 mW @ 1.095 GHz, 1.3V

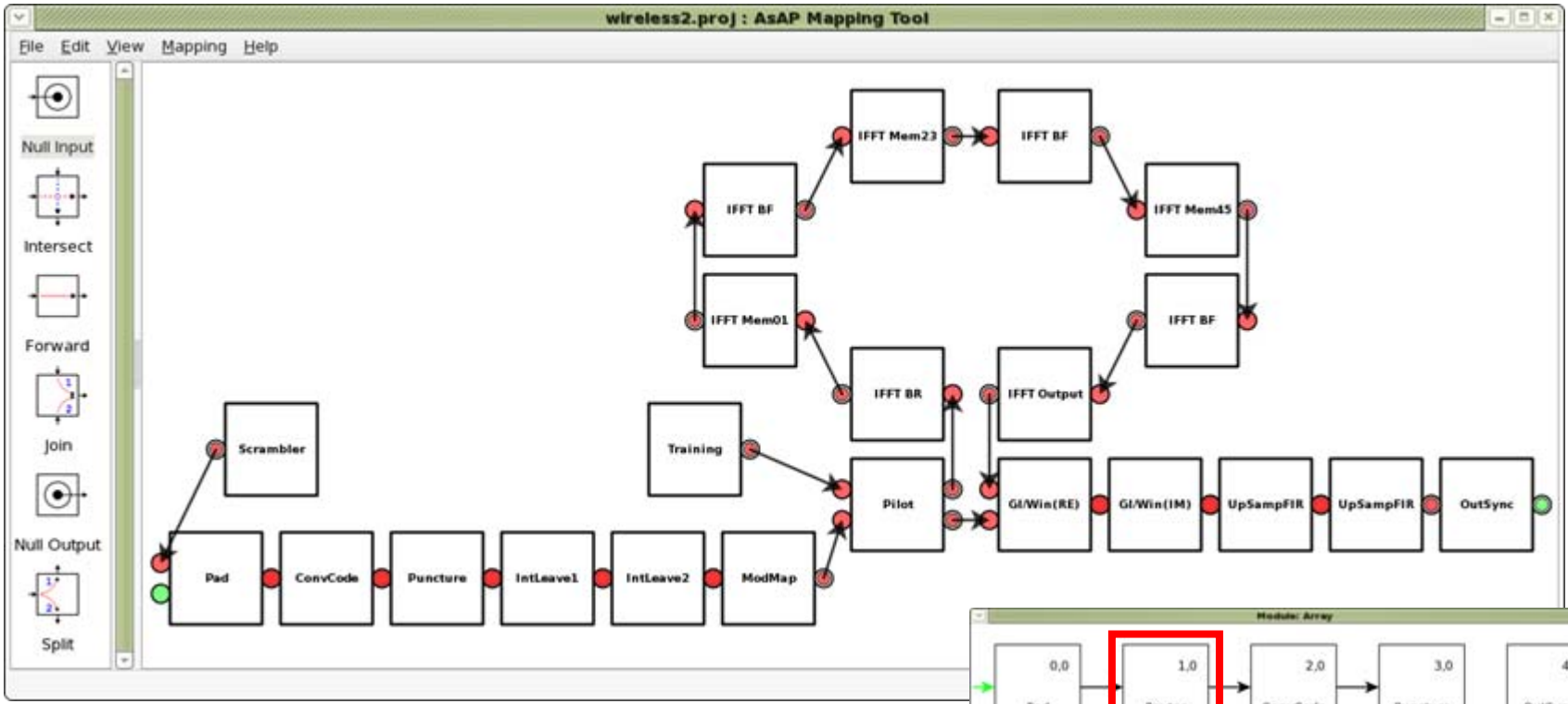


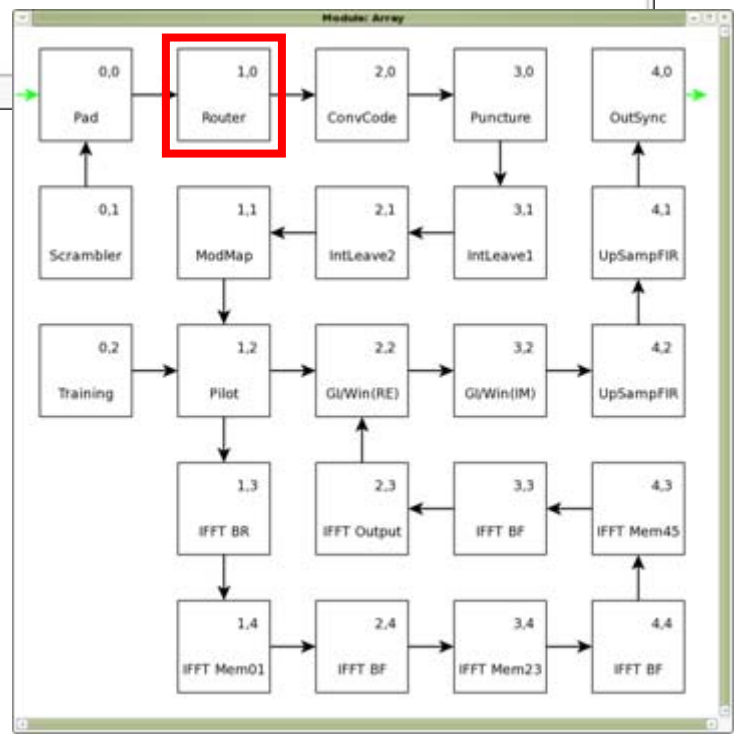
Programming

- Write C program(s) for each task
  - AsAP C compiler
- Connect programs with an arbitrary graph
- Auto-mapping tool
  - Maps to 2-D processor array
  - Adds routing processors if needed
  - Various optimization criteria









- 802.11a/11g transmitter constrained to a 5x5 array
- Router automatically added in (1,0)

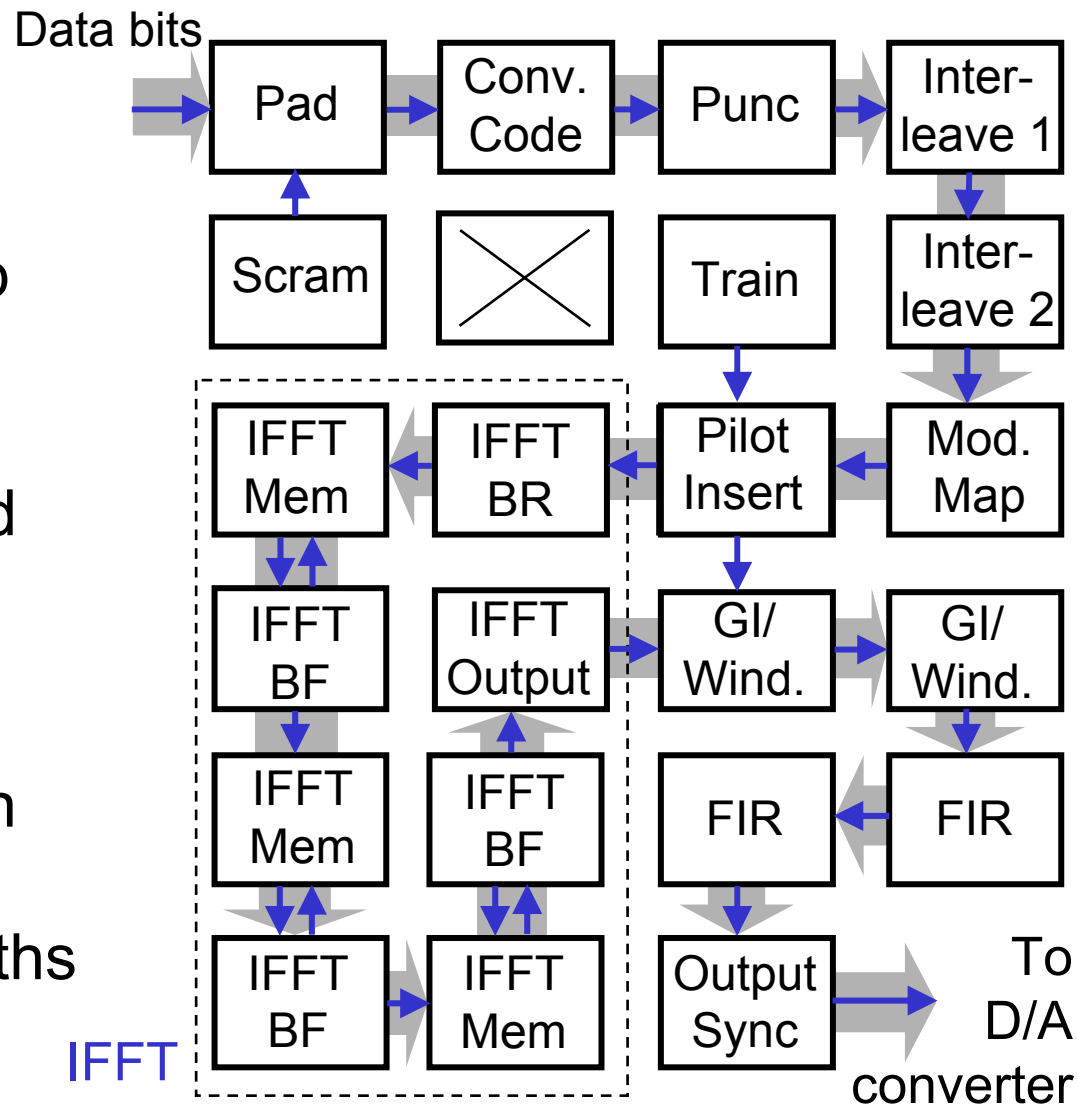
Example Tasks and Applications Completed

- FIR filters (~100)
- Convolution
- Sorting (bubble, merge)
- Division
- Square root
- CORDIC sin, cos, arcsin, arccos, arctan
- Natural log
- Exponential e^x
- Pseudo random number generation
- CRC calculation
- Matrix multiplication
- Huffman encoder
- 8-point Discrete Cosine Transforms
- 8×8 2-D DCT (several)
- Fast Fourier Transforms (FFTs) of length 32-1024
- Full $k = 7$ viterbi decoder

802.11a/802.11g WiFi Wireless LAN Transmitter (no special purpose procs)

- 22 processors
- Fully compliant with standard
- Fully functional on chip
- 407 mW @ 300 MHz 30% of 54 Mb/s
- Code unscheduled and lightly optimized
- 5x - 10x performance and 35x – 75x lower energy dissipation than 8-way VLIW TI C62x
- Programmed in 3 months

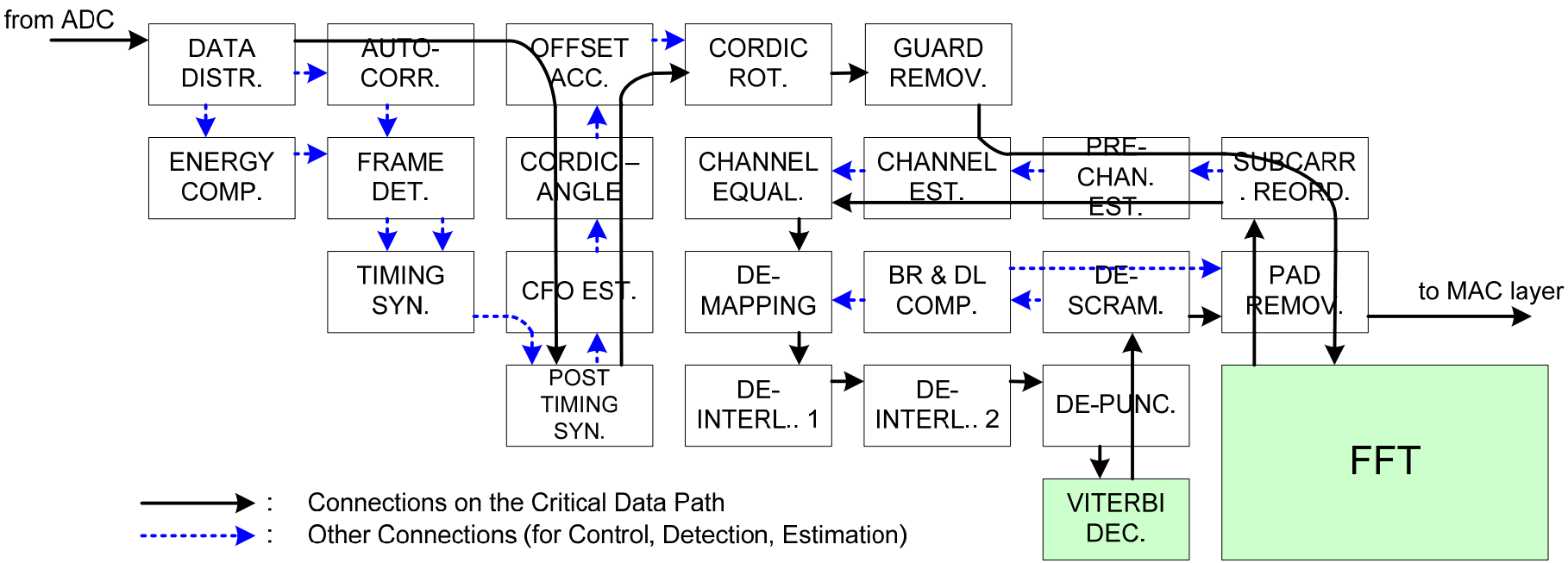
[SIPS 04; ICC 02]



Complete 802.11a Baseband Receiver

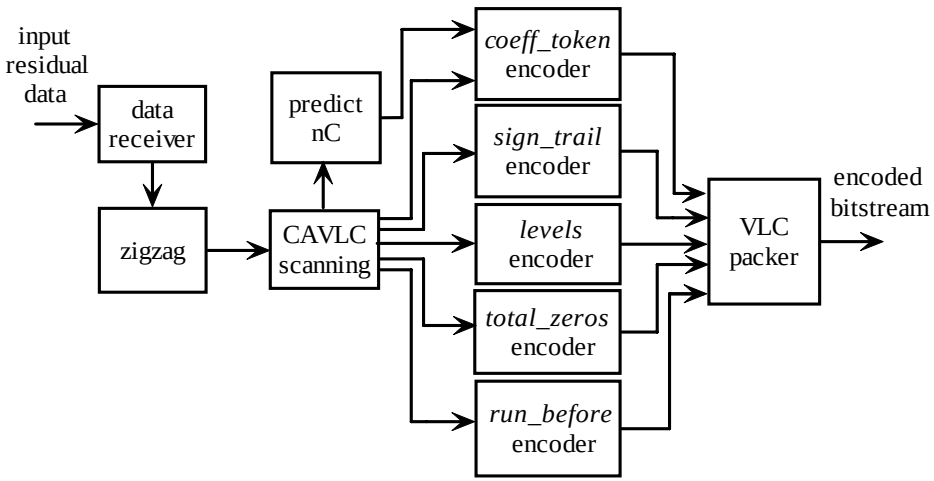
- 22+2 processors
- Fully compliant
- 54 Mbps throughput
  75 mW @ 610 MHz

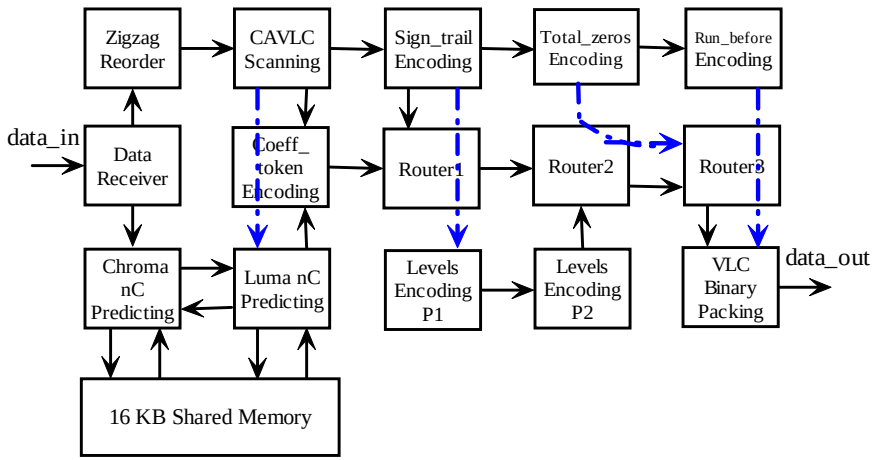
- 2-6x faster than specialized Software Defined Radio and general DSP processors when scaled to 65nm



Complete H.264 Context-Based Adaptive Length Coding (CAVLC) Encoder

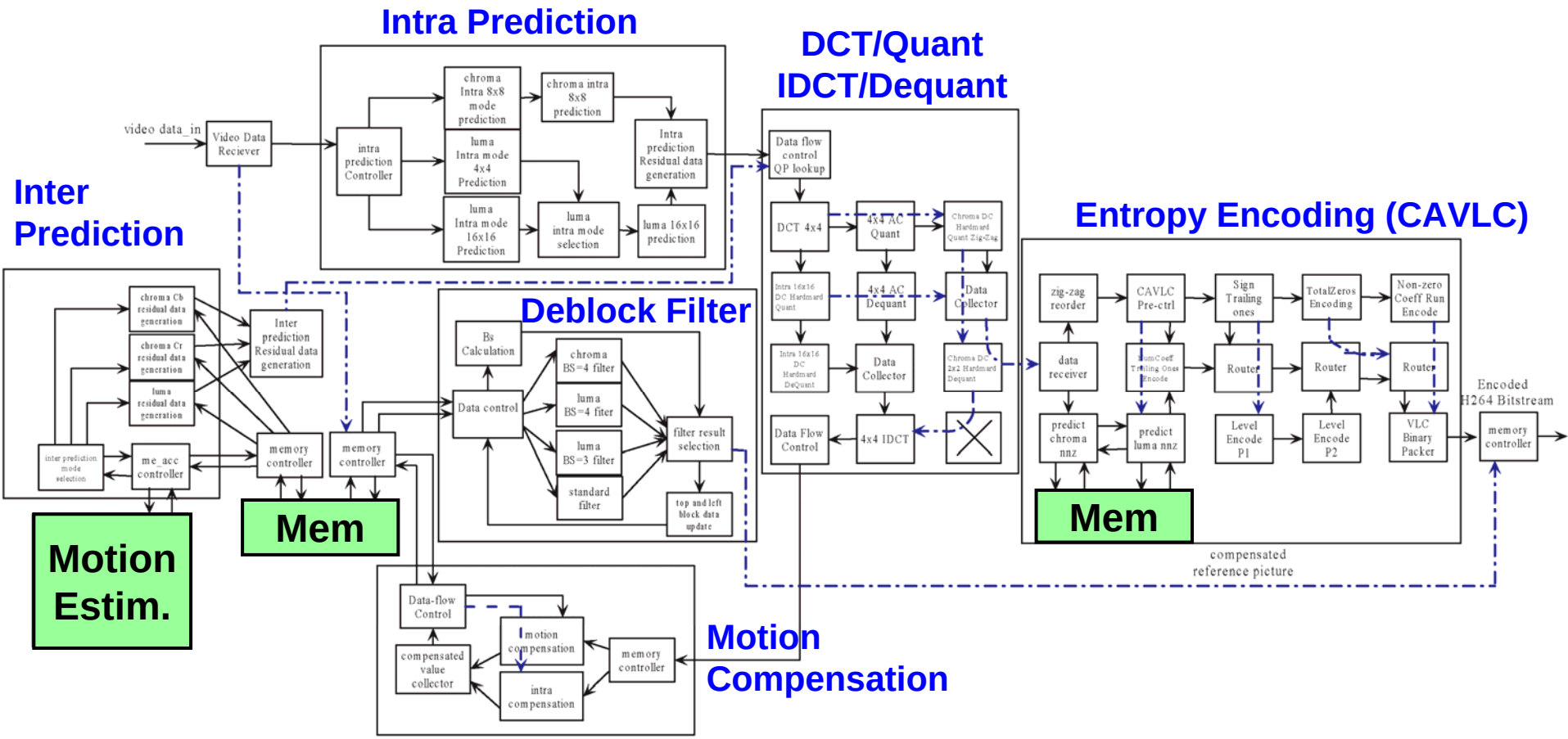
- 15 processors + (1) 16 KB memory
- Real time 720p HDTV @ 30 fps @ 1.07 GHz
- Compared to 3.6 GHz Pentium 4 HT
  - 4.9-6.8x faster and 20x smaller scaled circuit area
- Compared to scaled TI C641, C642, ADSP BF561
  - 1.0-6.1x faster and 6x smaller circuit area while computing a much more demanding frame sequence and *QP*=24 value





Complete H.264 Video Encoder

- Several 16 KB memories + motion estimation processor
- Off-chip frame buffers
- Estimated <100 processors for functional operation



Summary

- 164 homogenous cores
  - 1.2 GHz, 59 mW, 100% active @ 1.3 V
- Three dedicated-purpose processors
  - Typically 10x or higher speedup over AsAP processors
- Three 16 KB shared memories
- Efficient circuit-switched inter-processor interconnect and GALS clocking
  - Easily scalable
  - Variation-tolerant
  - Small area high-rate long-distance communication

Acknowledgments

- Funding and Support
  - SRC CSR Grant 1659.001
  - SRC GRC Grant 1598.001
  - ST Microelectronics
  - NSF CAREER award
  - NSF Grant No. 430090
  - Intel Corporation
  - UC Micro
  - Intellasys
  - SEM
  - UCD Faculty Research Grant
  - MOSIS
  - Artisan
- Special Thanks
  - R. Krishnamurthy, M. Anders, S. Mathew, P. Kudva, J.-P. Schoellkopf, Patrick Cogez, K. Torki, S. Muroor, W. Li