



1TOPS/W

Software Programmable Media Processor

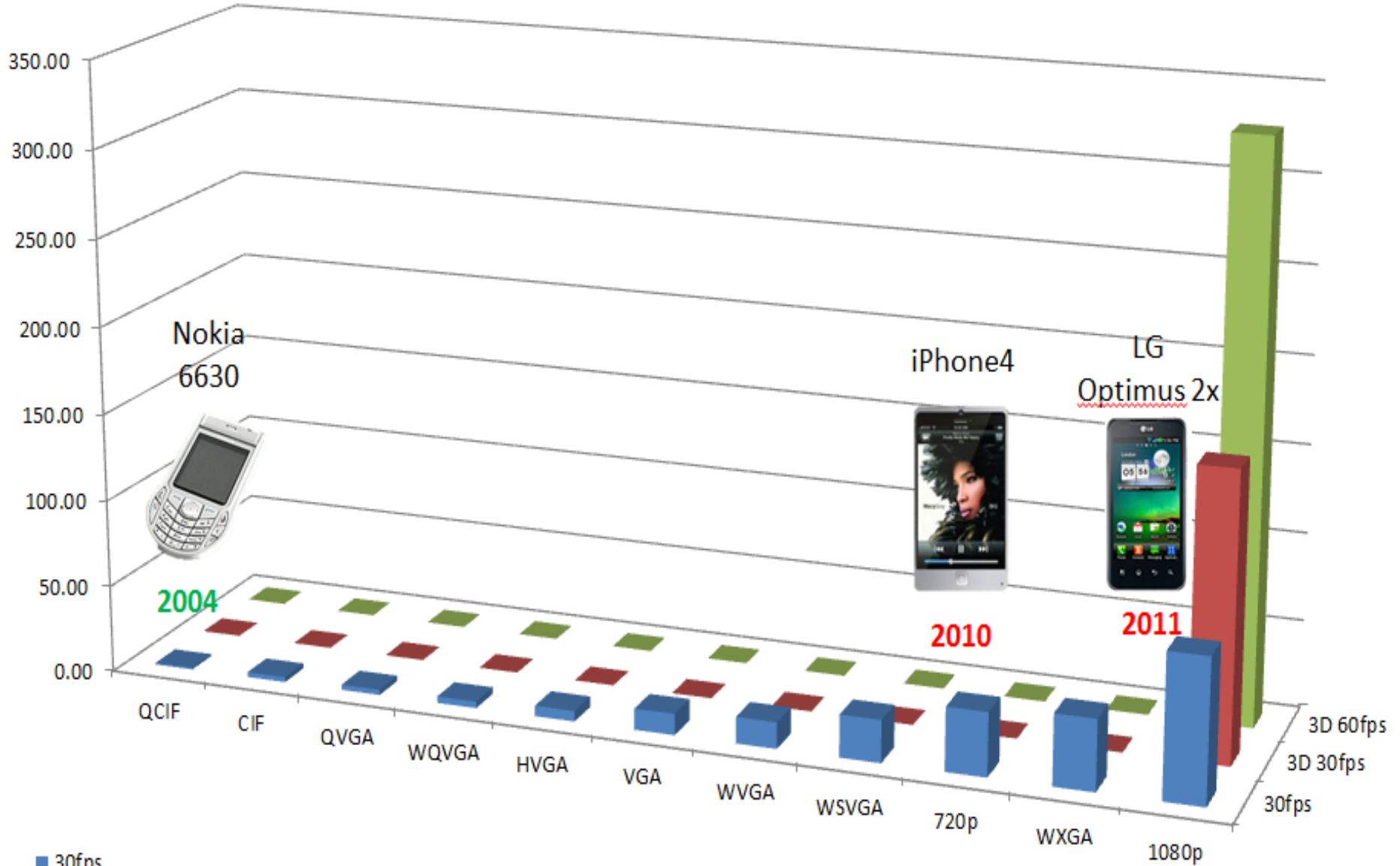
David Moloney, CTO, Movidius

19 August 2011

Movidius Background

- Started in 2005 looking at mobile gaming acceleration
 - Decided on multicore design to allow software derivatives and meet OPS/W/\$ target
 - Existing processors poor cost/performance match for target workloads
 - Developed SHAVE vector processor with HW support for sparse data-structures (Matrix-Vector)
 - Expanded ISA to support C-complier
- Talked to mobile phone customers in 2007
 - Turned out their real problem was video
 - Back to the drawing-board!
- Initial 65nm Silicon & all IP on founder & angel funding
 - Allowed us to close A-round in October 2008!
- 65nm Myriad MM SoC in mass-production
 - Next generation 28nm SoC H1/2012 with 10x Perf/W

Mobile Video Processing Workload



■ 30fps

20/Apr/2011

Movidius SHAVE Processor

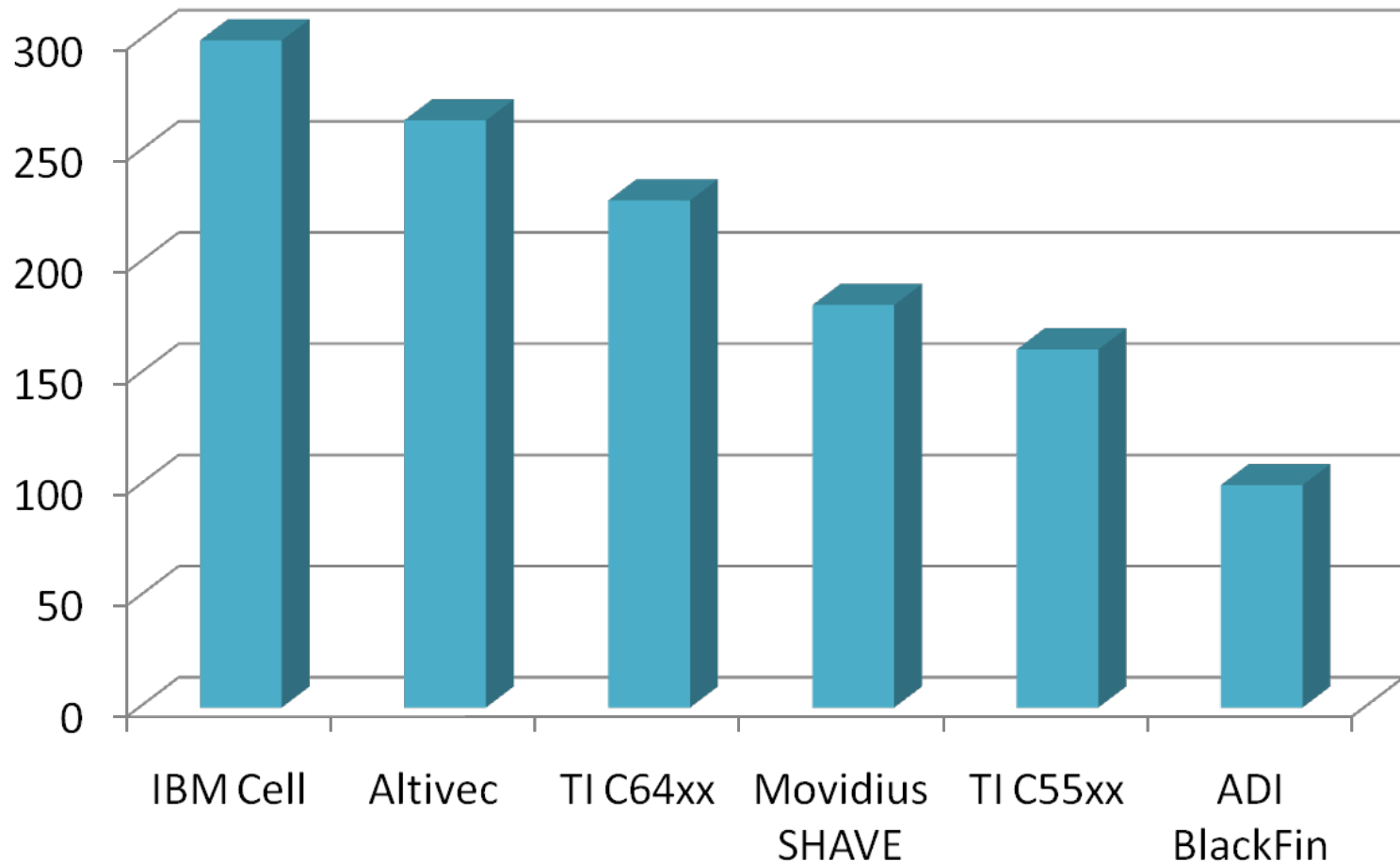
- Steaming Hybrid Architecture Vector Engine
 - Hybrid of RISC, DSP, VLIW & GPU architectural features
 - 128-bit vector arithmetic: 8/16/32-bit INT & fp16/fp32
- Unique proprietary architecture
 - Tailored to streaming workloads and architected for outstanding OPS/mW/\$ performance
- Excellent Graphics and matrix mathematics support
 - HW texture unit for good graphics performance
 - Predicated execution to eliminate branches
 - Compiler-friendly architecture
 - HW support for compressed data-structures (ex. matrices)

SHAVE Instruction-Set

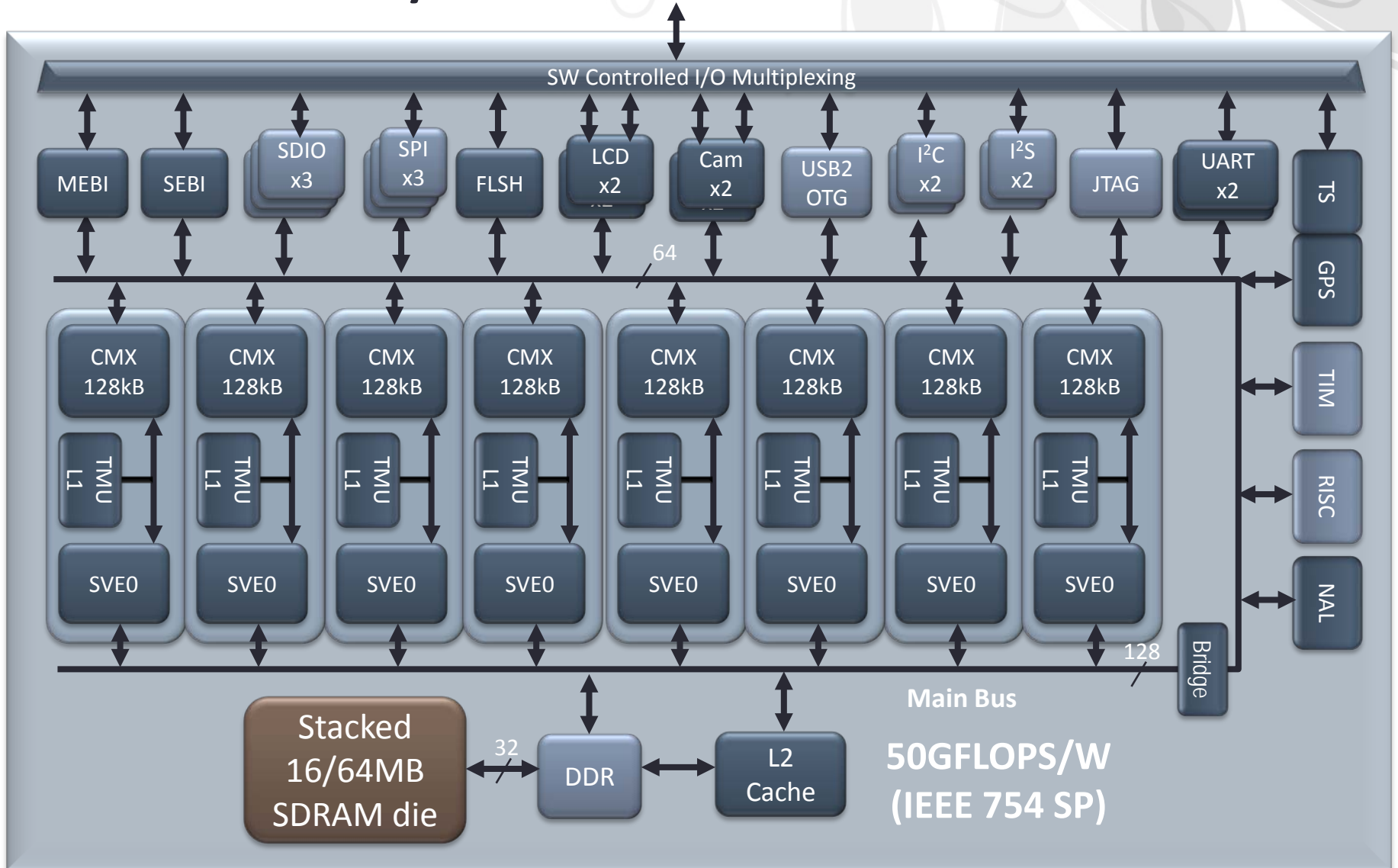
- RISC-style
 - Instruction predication
 - Extensive integer ISA
 - Excellent C-compiler support
- DSP-style
 - Zero overhead looping
 - Modulo addressing
 - Transparent DMA modes
 - FFT, Viterbi, and other DSP operation support
 - Parallel comparisons
- VLIW-style
 - Parallel functional units controlled by VLIW instr.
 - 8/16/32-bit x 1-4 SIMD INT
- GPU-style
 - Streaming operations
 - Floating-point operations (fp16/fp32 IEEE-compliant)
 - Texture-Management Unit and L1 Cache

SHAVE ISA Richness

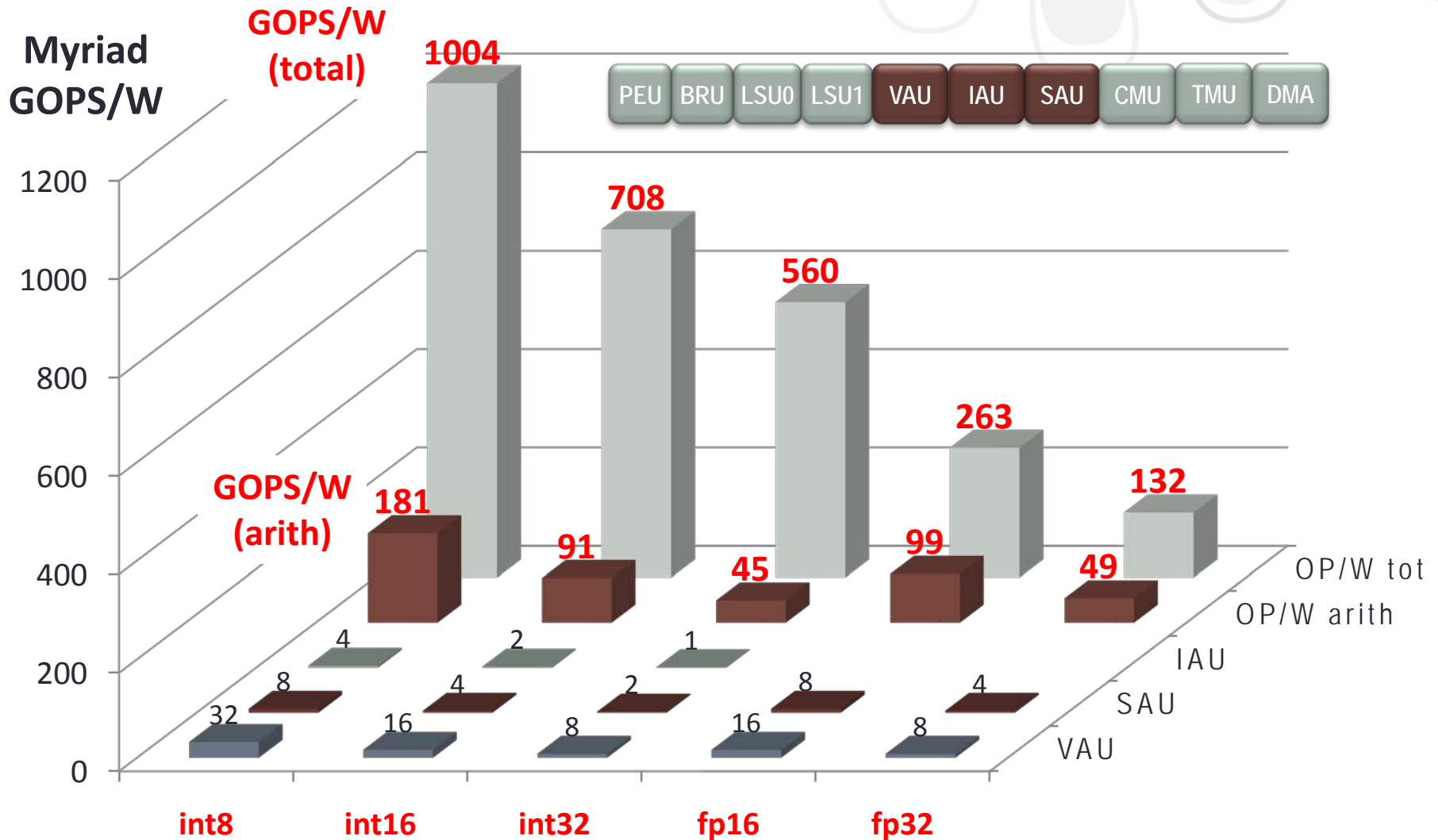
instructions



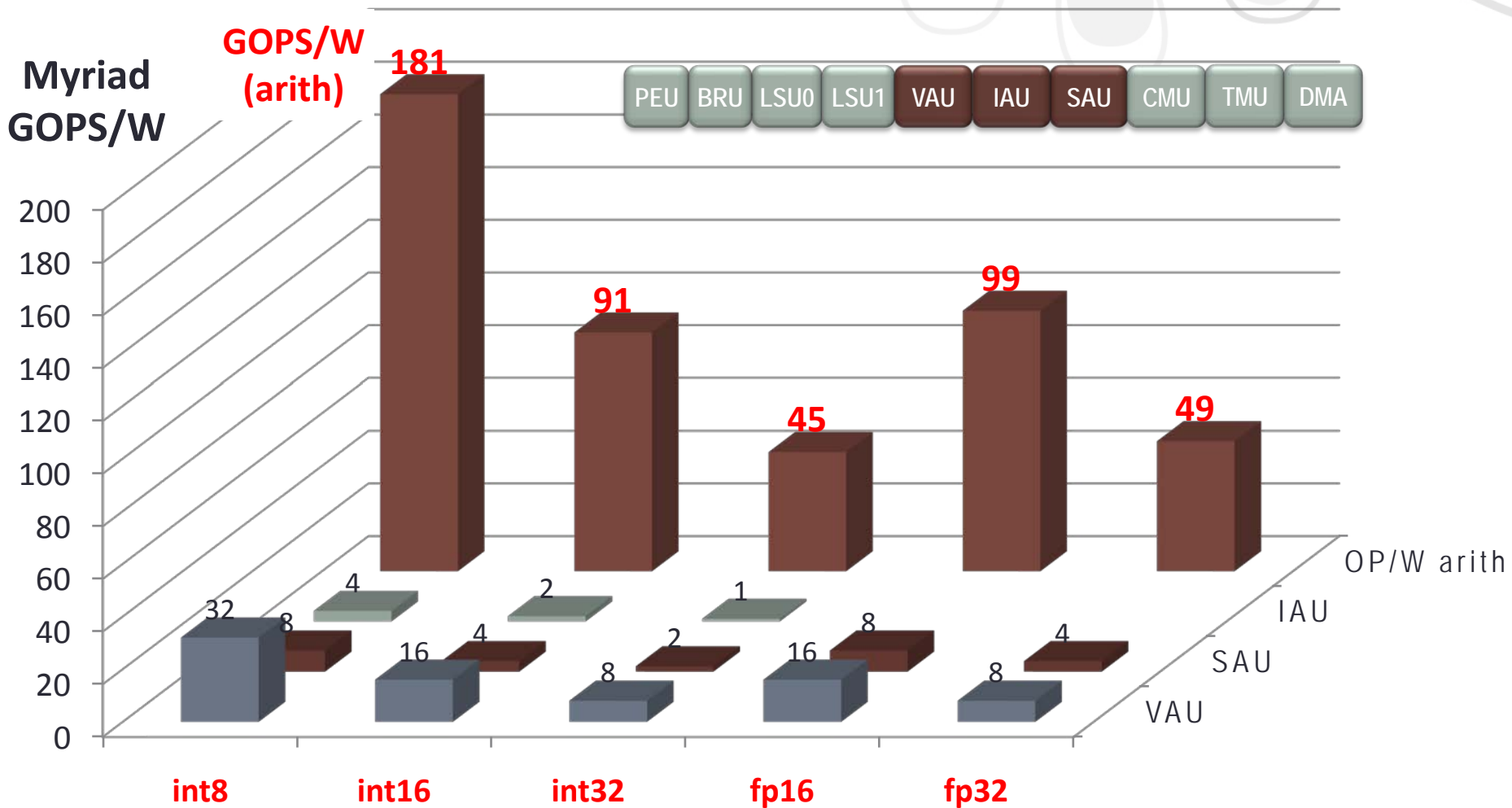
Myriad Silicon Platform



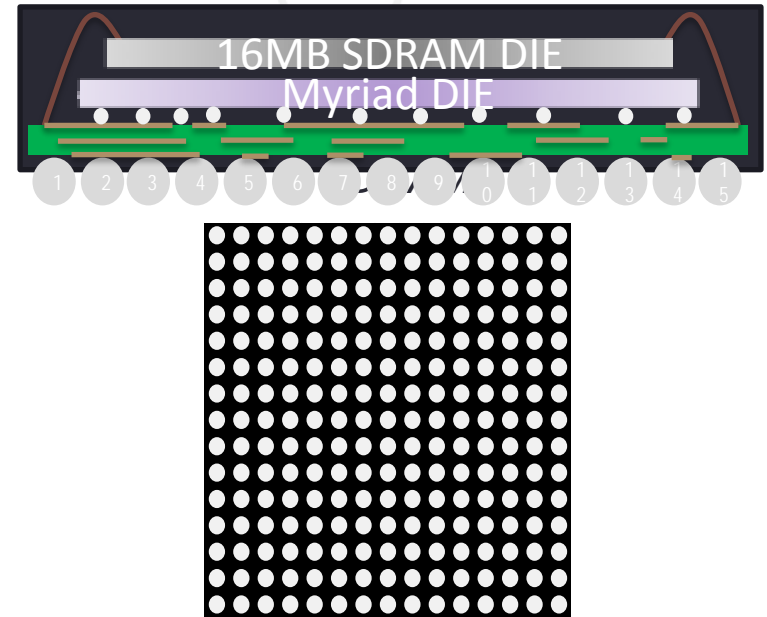
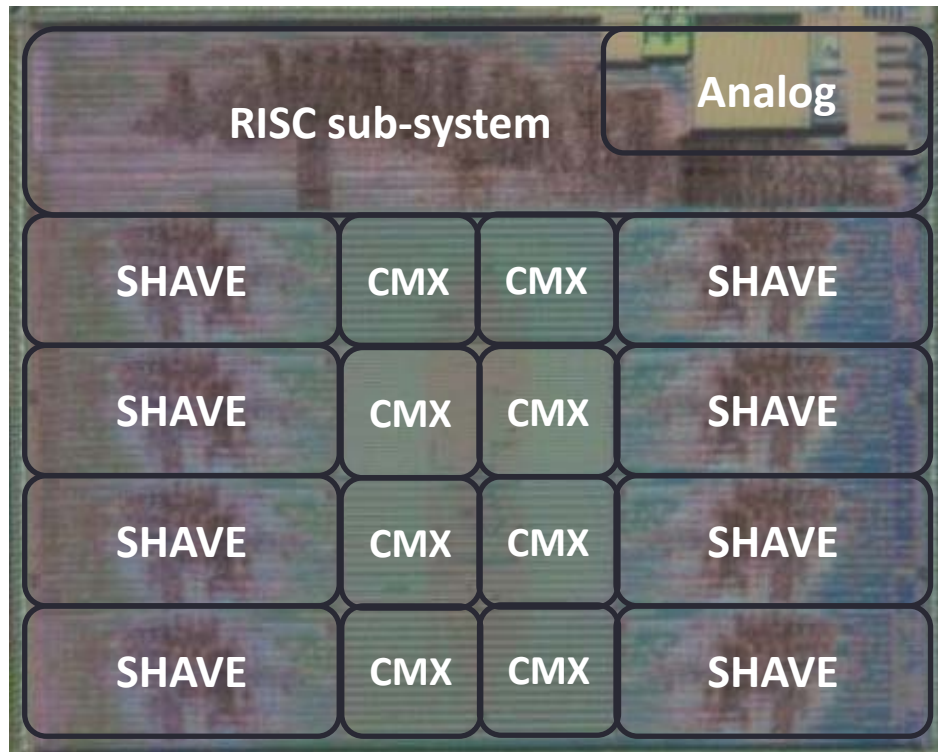
Myriad GOPS/Watt (Total)



Myriad GOPS/Watt (Arithmetic)



Myriad 65nm CMOS LP Die



	Author	Year	FLOPS/core	Cores	GFLOPS	W	GFLOPS/W
Myriad	Movidius	2011	12	8	17.28	0.35	49.4
(1	KAIST	2011			5.8	0.28	21.1
(2	Intel	2007		80	1000	98.00	10.2
(4	Adapteva	2010	2	16	24.96	1.00	25.0

Technology - Platform Approach



Applications

Software Modules

Silicon Platform

Foundation Technology

Products

3D Capture



Video Edit



3D Video



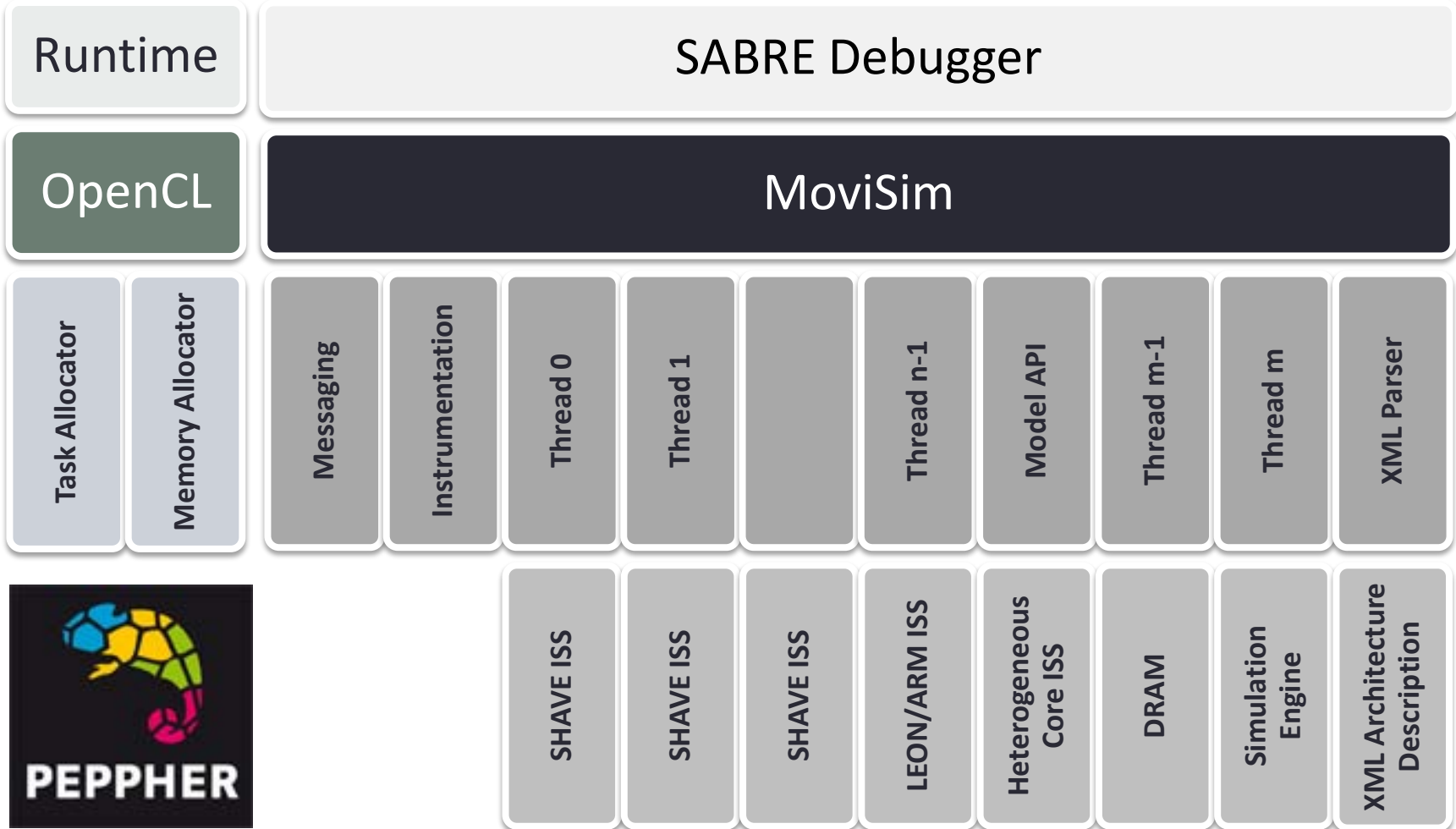
Anaglyph-3D



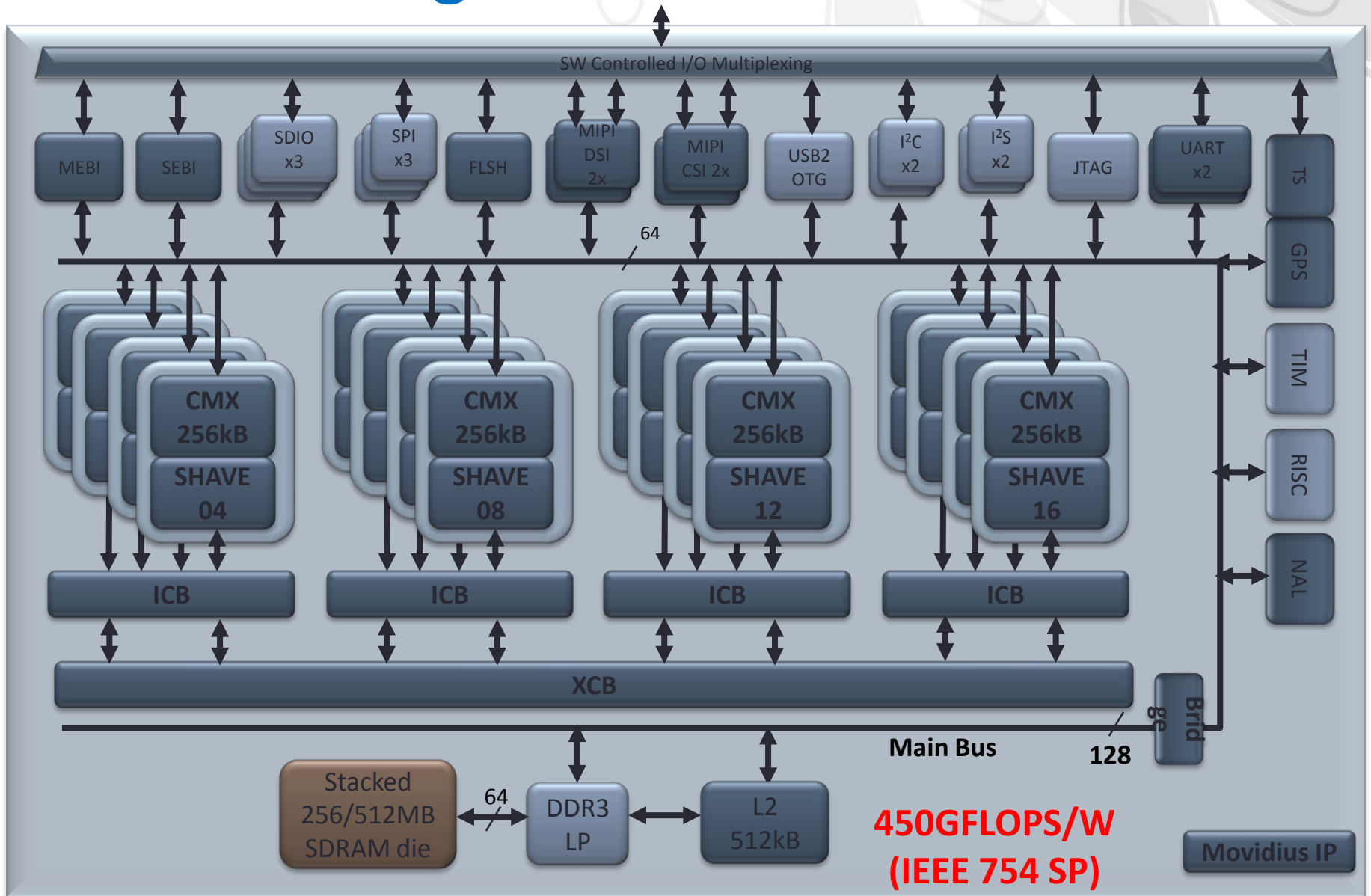
Myriad Example Applications



MoviSim ISS Architecture



Fragrak 28nm Platform





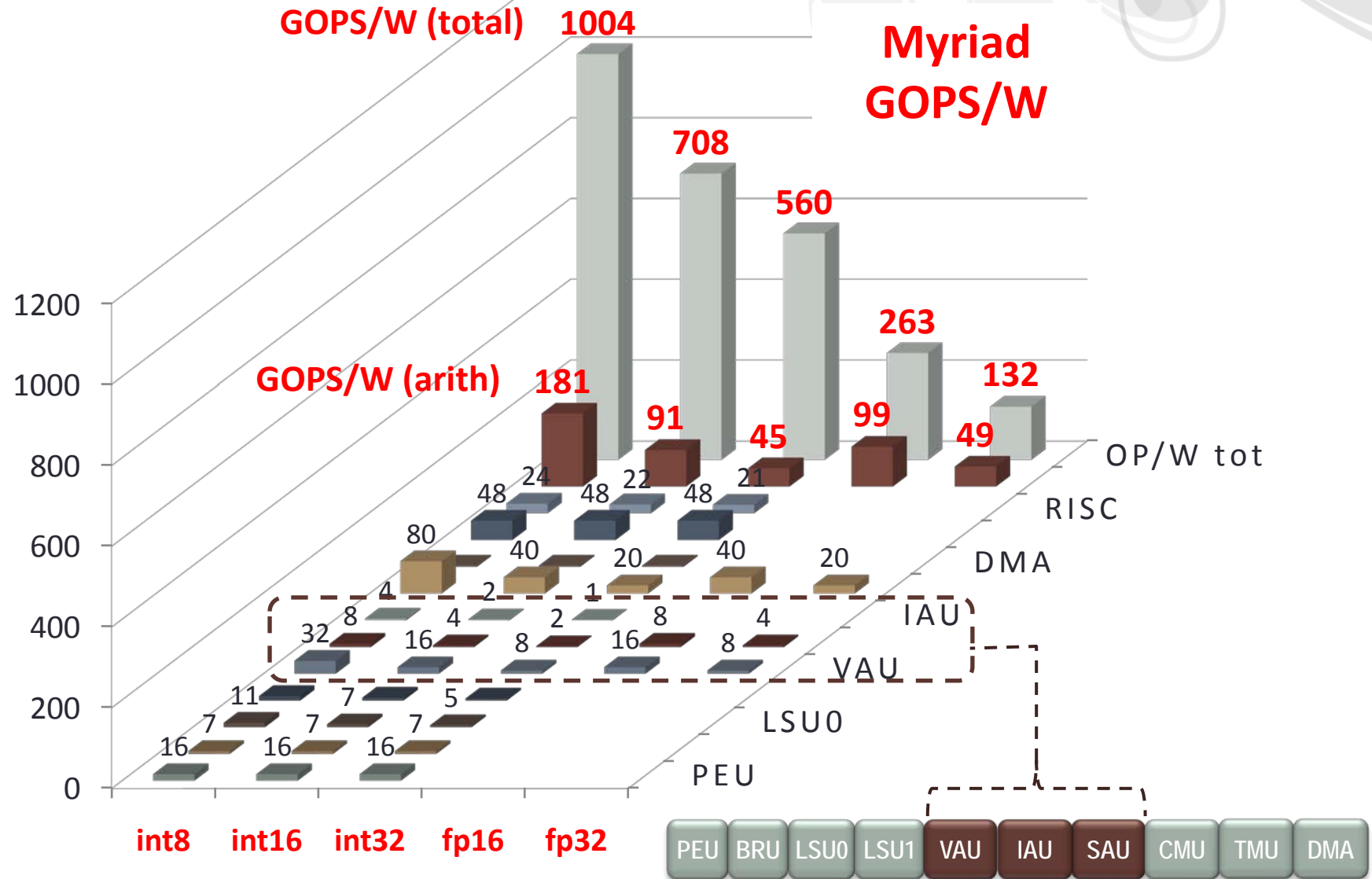
Any questions?

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°248481 (PEPPHER Project, www.peppher.eu)

Abstract

- The rationale and architecture behind a new software programmable multimedia coprocessor for mobile devices is outlined.
- The focus of the architecture is on power-efficient operation, allowing functions which are traditionally implemented in fixed-function hardware to be implemented competitively in software.
- For instance the sustained single-precision IEEE 754 rate is 50GFLOPS/W allowing existing applications to be ported with great ease. The device supports 8, 16, 32 and some 64-bit integer operations as well as fp16 (OpenEXR) and fp32 arithmetic and is capable of an aggregate 1 TOPS/W maximum 8-bit equivalent operations in a low-cost plastic BGA package with integrated 16 or 64MB SDRAM.
- New architectural features such as support for random-accessible sparse data-structures are implemented for the first time improving memory utilization and bandwidth efficiency. Power efficiency is paramount and the device contains a total of 11 power-islands with 8 dedicated to each of the integrated SHAVE processors, allowing very fine-grained power control.
- Comparisons to previous work based on 65nm silicon and applications are shown to illustrate the power of the device.

Myriad GOPS/Watt (Total/Arithmetic)

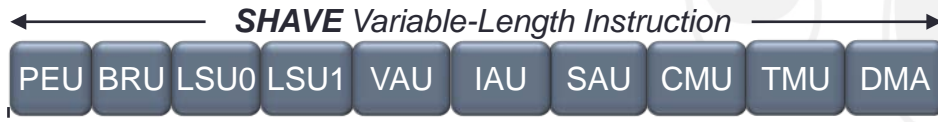


16x2x800MHz =
25.6GB/Sec

SHAVE Processor 28nm (Fragrak)

800MHz

Intra-Cluster
Bus (ICB)



8x2x800MHz =
12.8GB/Sec

256kB
Per
SHAVE

256kB
SRAM CMX

Xtra-Cluster
Bus (XCB)

16x2x800MHz
25.6GB/Sec

SHAVE
Processor

PEU BRU DCU

CMU

TMU

IRF
32x32

SRF
32x32

VRF
32x128

16k
L1

IAU

SAU

VAU

LSU0

LSU1

DCU

Decoded
Instructions

512
kB

L2
Cache
512kB
2-way

256 -
512MB
SDRAM

256 -
512MB
SDRAM
Die

LP
DDR3
Cont.

128-bit AXI Bus

16x800MHz
12.8GB/Sec

4x17x800MHz
54.4GB/Sec

4x12x800MHz
38.4GB/Sec

16x12x800MHz
76.8GB/Sec

16x2x800MHz
25.6GB/Sec

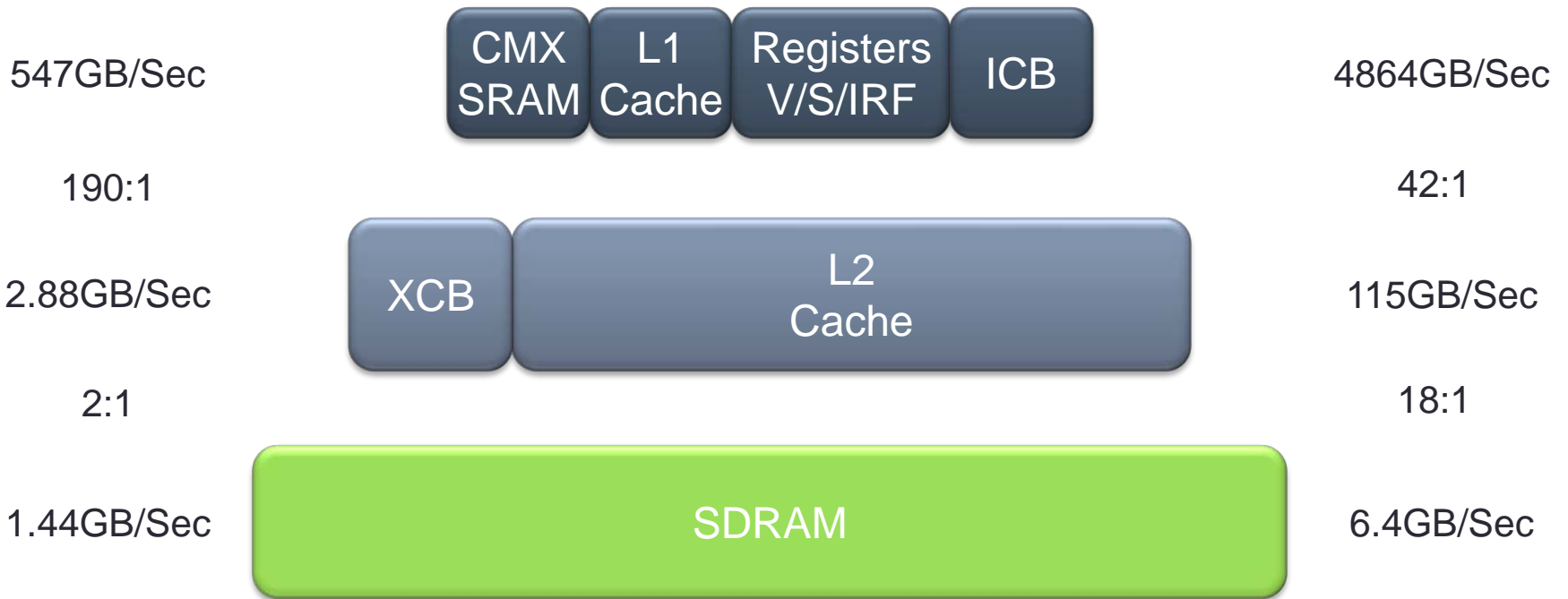
8x2x800MHz
12.8GB/Sec

Fragrak

BW Hierarchy

*Myriad
65nm*

*Fragrak
28nm*



Bottom-Line - **Very High Sustainable Performance**

BW Hierarchy (Detail)

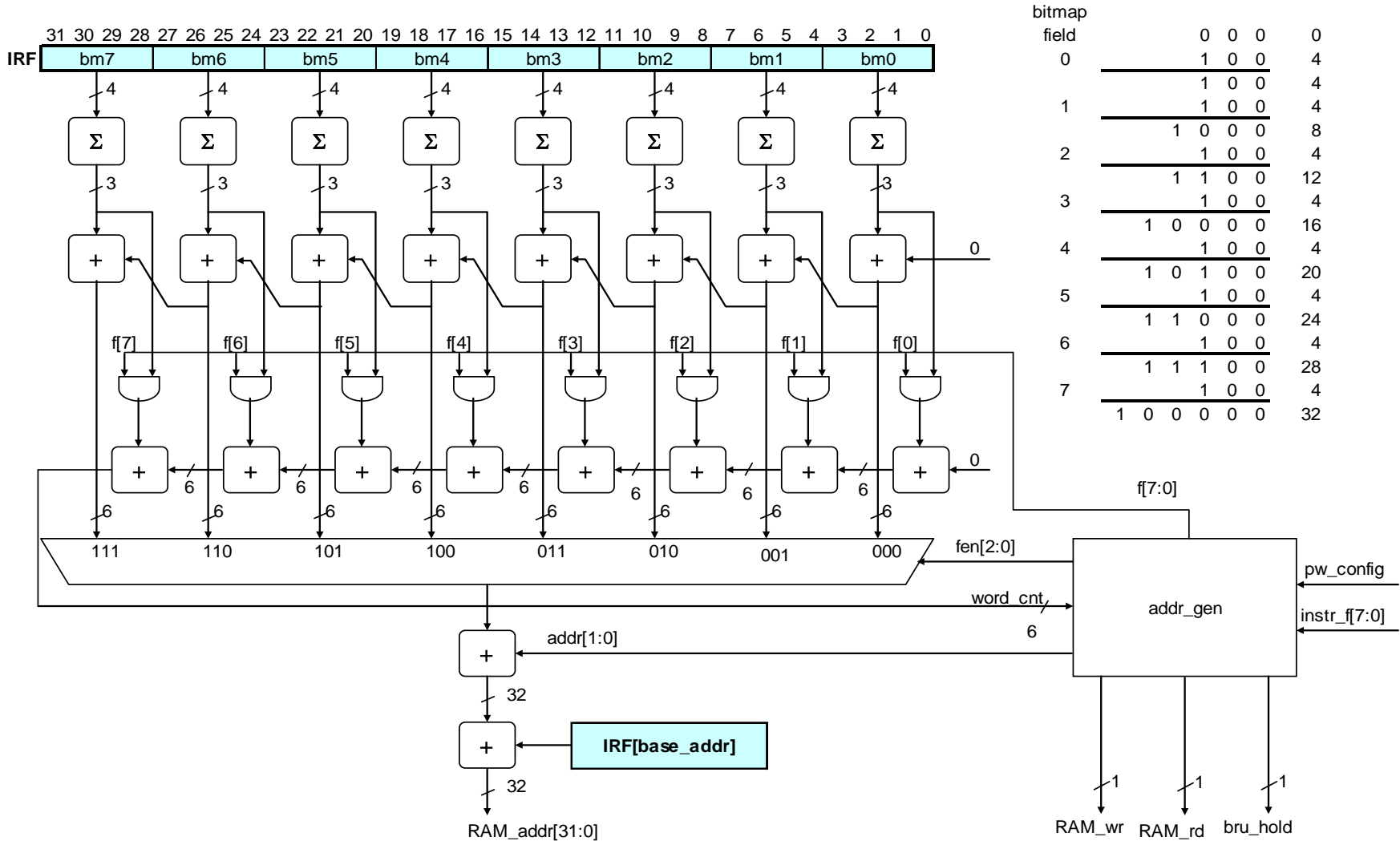
Myriad 65nm

	VRF	SRF	IRF	LSU	IDC	L1	ISB	L2	SDRAM
Clk	180	180	180	180	180	180	180	180	180
Bytes	16	4	4	8	16	8	16	16	4
Ports	12	12	17	2	1	1	2	1	2
BW	34.56	8.64	12.24	2.88	2.88	1.44	5.76	2.88	1.44
#SHAVES	8	8	8	8	8	8	8		
Total BW	276.48	69.12	97.92	23.04	23.04	11.52	46.08		
	547.2							2.88	1.44
								190	2

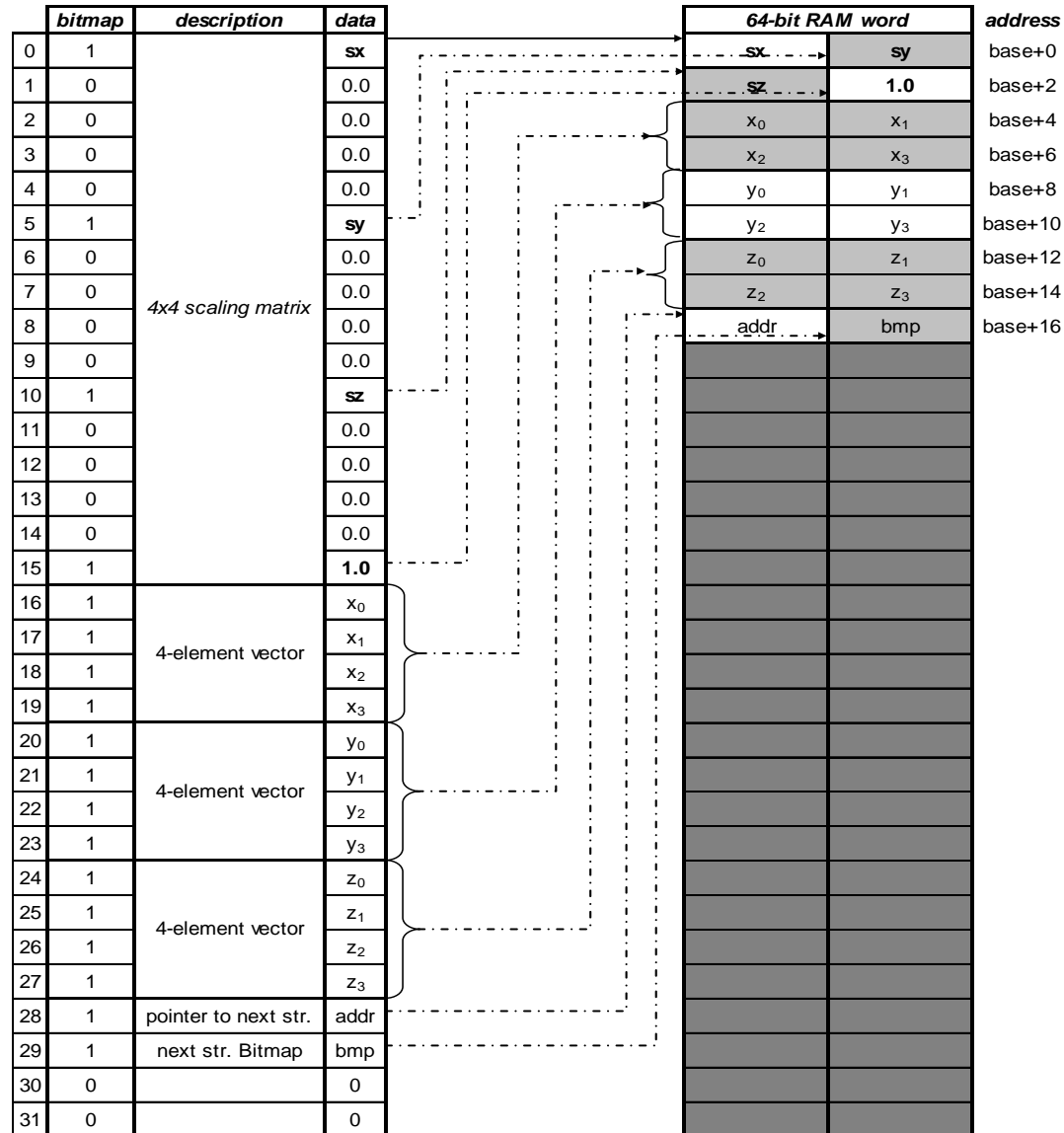
Fragrak 28nm

	VRF	SRF	IRF	LSU	IDC	L1	ICB	L2	XCB	SDRAM
Clk	800	800	800	800	800	800	800	800	800	800
Bytes	16	4	4	8	16	8	16	16	16	4
Ports	12	12	17	2	1	1	2	1	8	2
BW	153.6	38.4	54.4	12.8	12.8	6.4	25.6	12.8	102.4	6.4
#SHAVES	16	16	16	16	16	16	16			
Total BW	2457.6	614.4	870.4	204.8	204.8	102.4	409.6			
	4864							115.2	6.4	
								42.22222222	18	
	8.888888889							40	4.444444444	

LSU HW Sparse-Data Support



Sparse Data-Structure Example



References

- 1) H-E. Kim, J-S. Yoon, K-D. Hwang, Y-J. Kim, J-S. Park, L-S. Kim, "A 275mw heterogeneous Multimedia processor for ic-Stacking on Si-interposer" Proc. ISSCC 2011
- 2) S.Vangal, J.Howard, G.Ruhl, S.Dighe, H.Wilson, J.Tschanz, D .Finan, P.Iyer,A. Singh, T.Jacob, S.Jain, S.Venkataraman, Y.Hoskote and N.Borkar, "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS", Proc. ISSCC 2007, pp.5-7
- 3) A. Olofsson, R. Trogan, O. Raikhman, "A 25 GFLOPS/Watt Software Programmable Floating Point Accelerator", HPEC 2010, 15-16 Sep 2010
- 4) C.Y. Park, N.I. Cho, "A fast algorithm for the conversion of DCT coefficients to H.264 transform coefficients", ICIP 2005 Proceedings, pp.664-7