SGRT: A Mobile GPU Architecture for Real-Time Ray Tracing

Won-Jong Lee[1], Youngsam Shin[1], Jaedon Lee[1],
Jin-Woo Kim[2], Jae-Ho Nah[3], Seokyoon Jung[1], Shihwa Lee[1],
Hyun-Sang Park[4], Tack-Don Han[2]

[1]SAMSUNG Advanced Institute of Technology
[2]Yonsei Univ., [3]UNC, [4]National Kongju Univ.

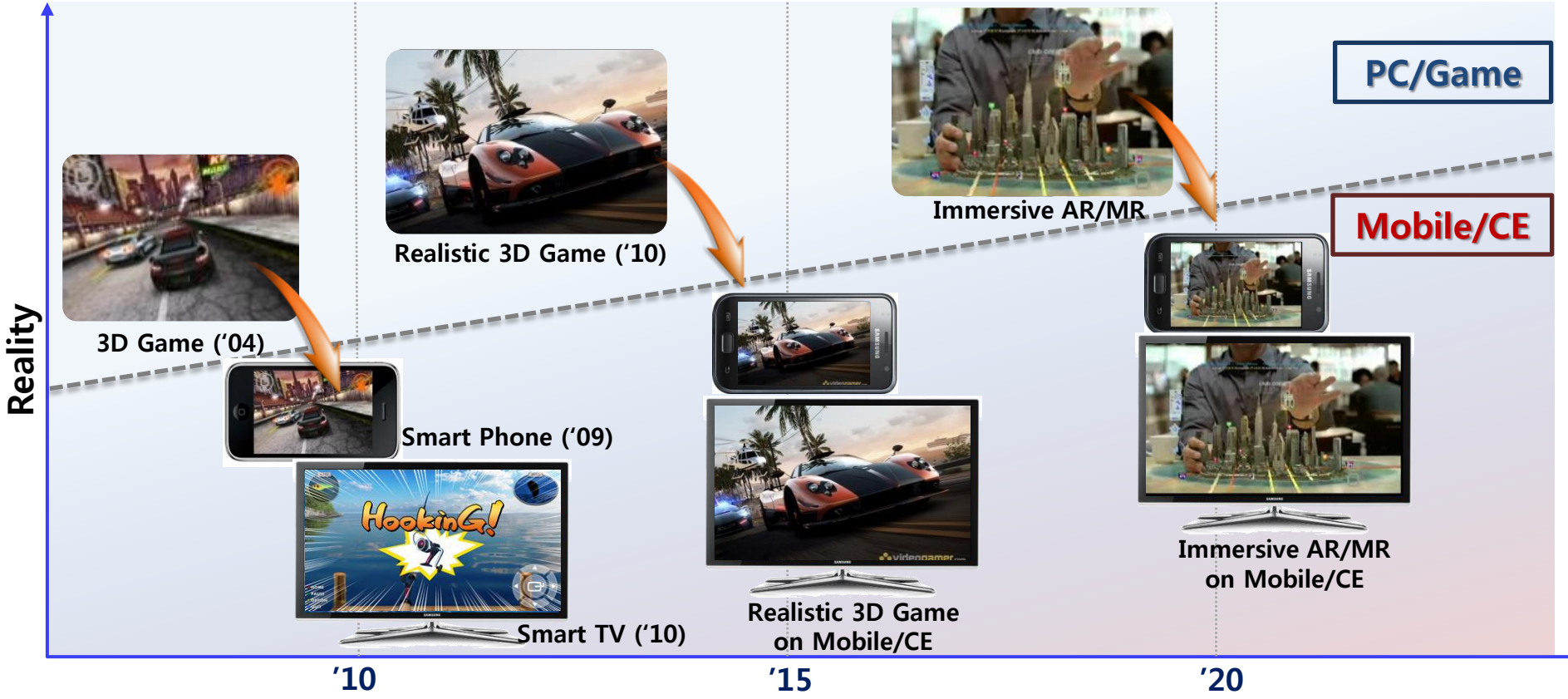
Outline

- **Introduction**
- **Related Work**
- **Proposed System Architecture**
  - Basic design decision
  - Dedicated hardware for T&I
  - Reconfigurable processor for RGS
- **Results and Analysis**
- **Conclusion**

Introduction

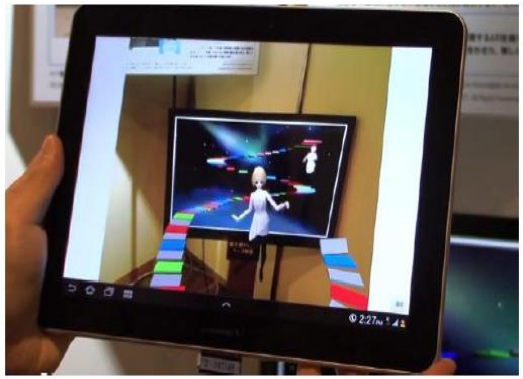
Graphics Trends

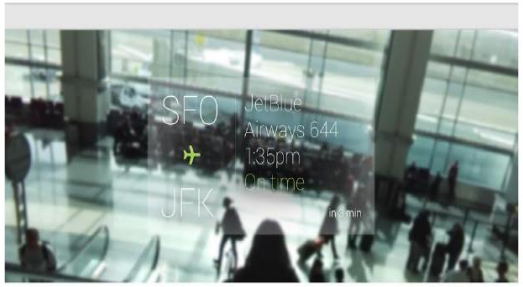
- Graphics is being important as increasing smart devices
- Evolving toward more realistic graphics
- Mobile graphics template earlier PC graphics (4~5 years)

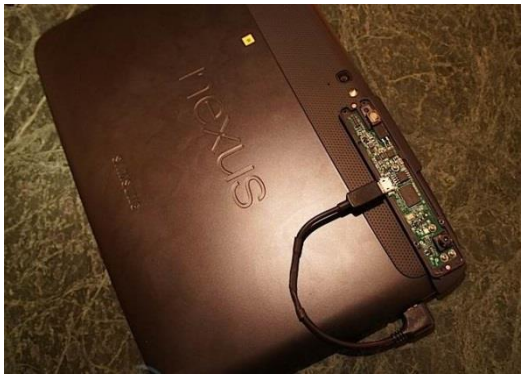


Future Mobile Graphics – Mixed Reality

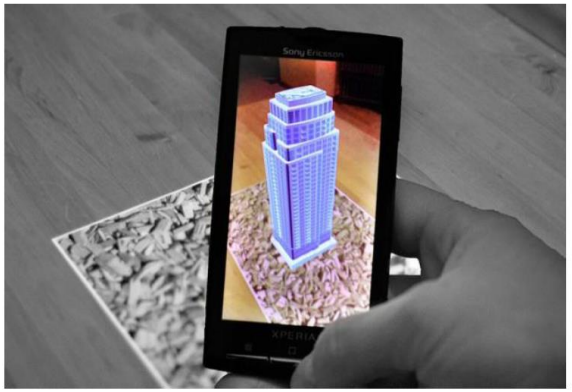
- Fusion of the physical world and a virtual world
- Estimated 6.6 M glasses-like devices in 2016 (IHS Research)
- Ray-traced objects will be naturally mixed with real-world objects and make the AR/MR application more immersive







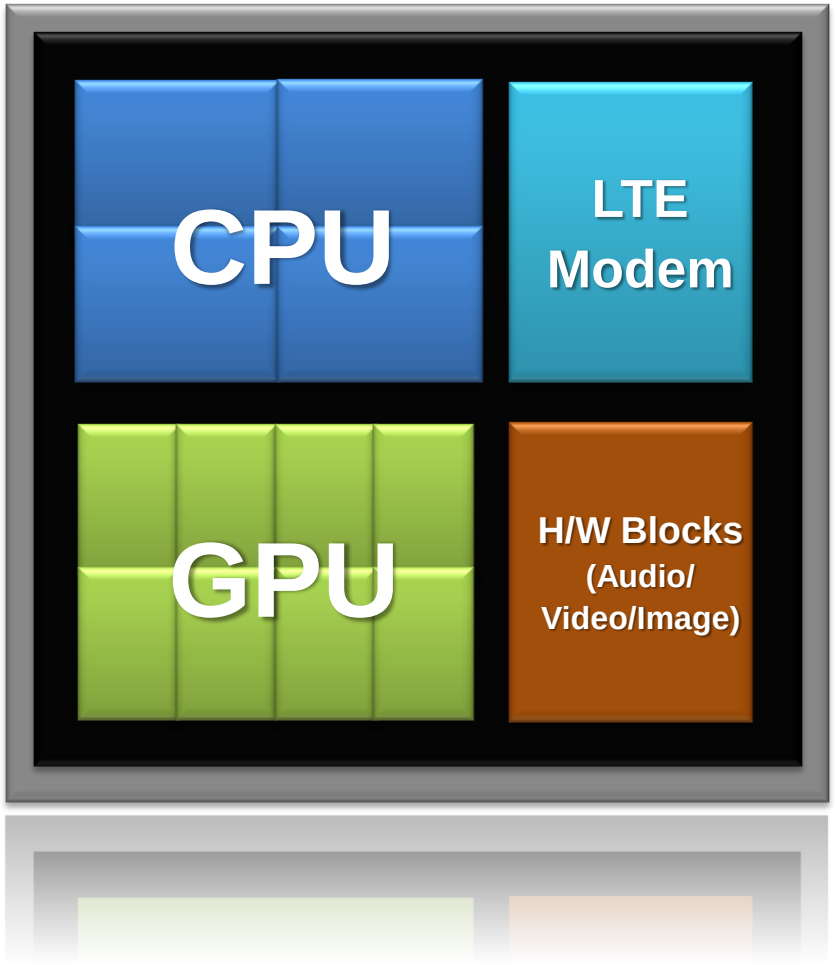
Capri depth camera attached to a mobile devices



Today's AR doesn't light the virtual objects to match the scene.

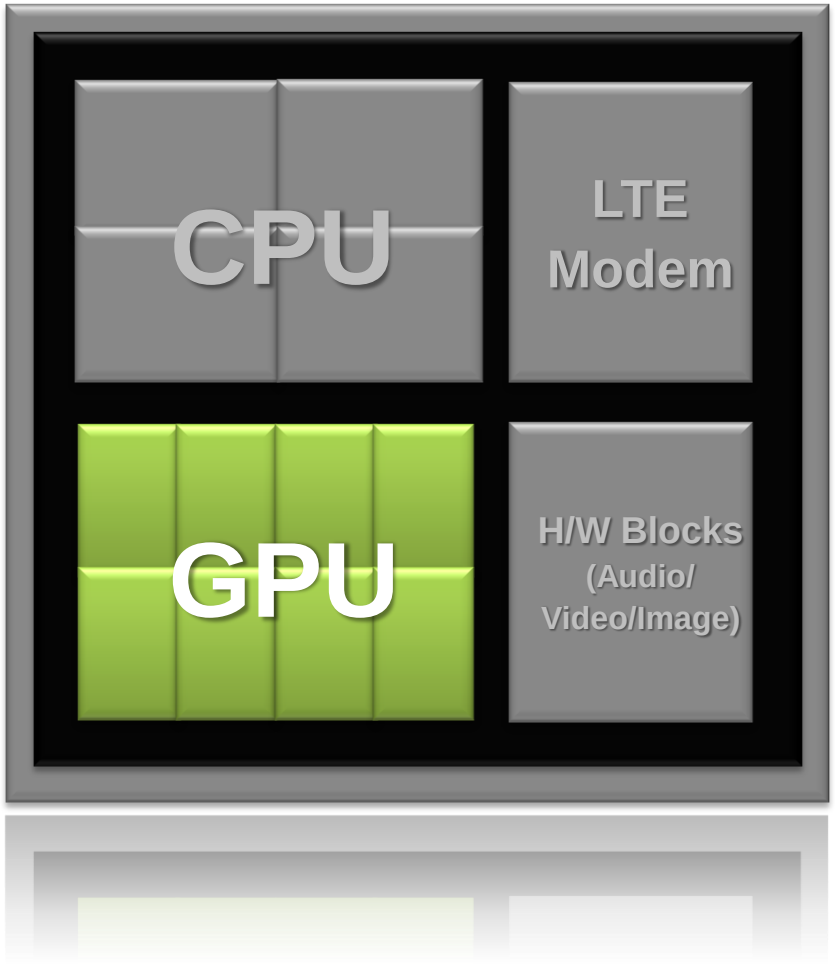
Naturally lit AR spheres in scene

Modern Mobile Computing Platform



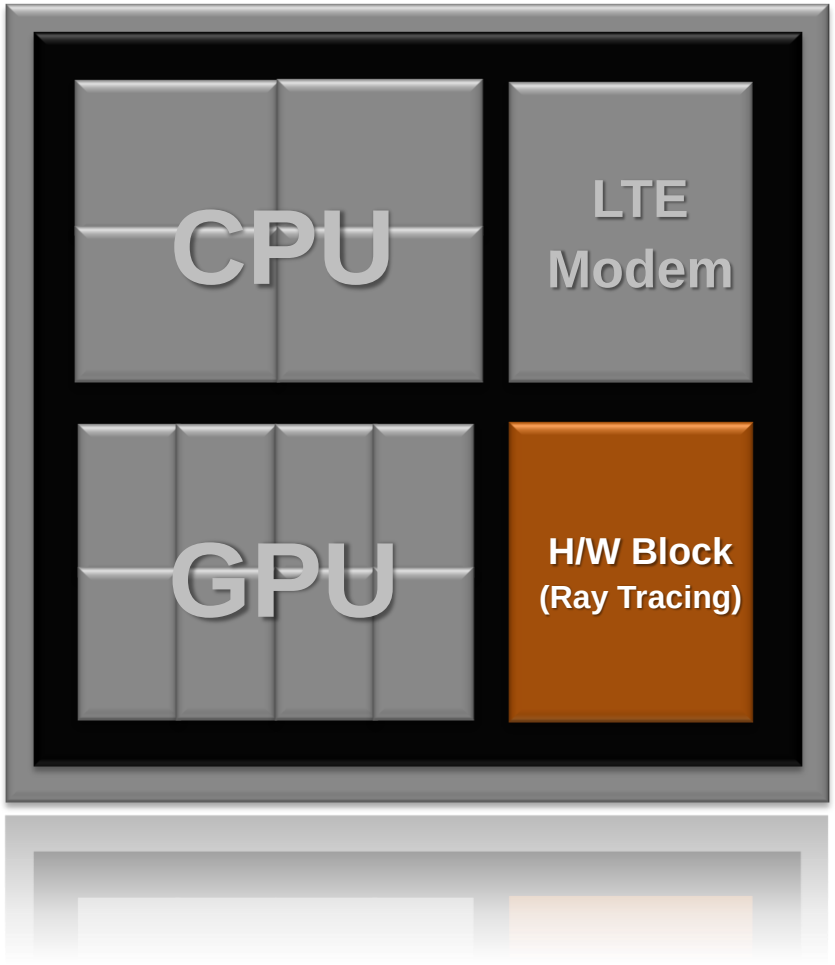
- **Multi-core CPU**
  - General program task
  - Fat-cores (highly sequential)
  - Multi-level memory hierarchies
- **Multi-core GPU**
  - Thin-cores (massively parallel)
  - Graphics, GPGPU
- **Fixed function H/W blocks**
  - H/W specialization for low-power multimedia processing

Ray Tracing on Mobile GPU?



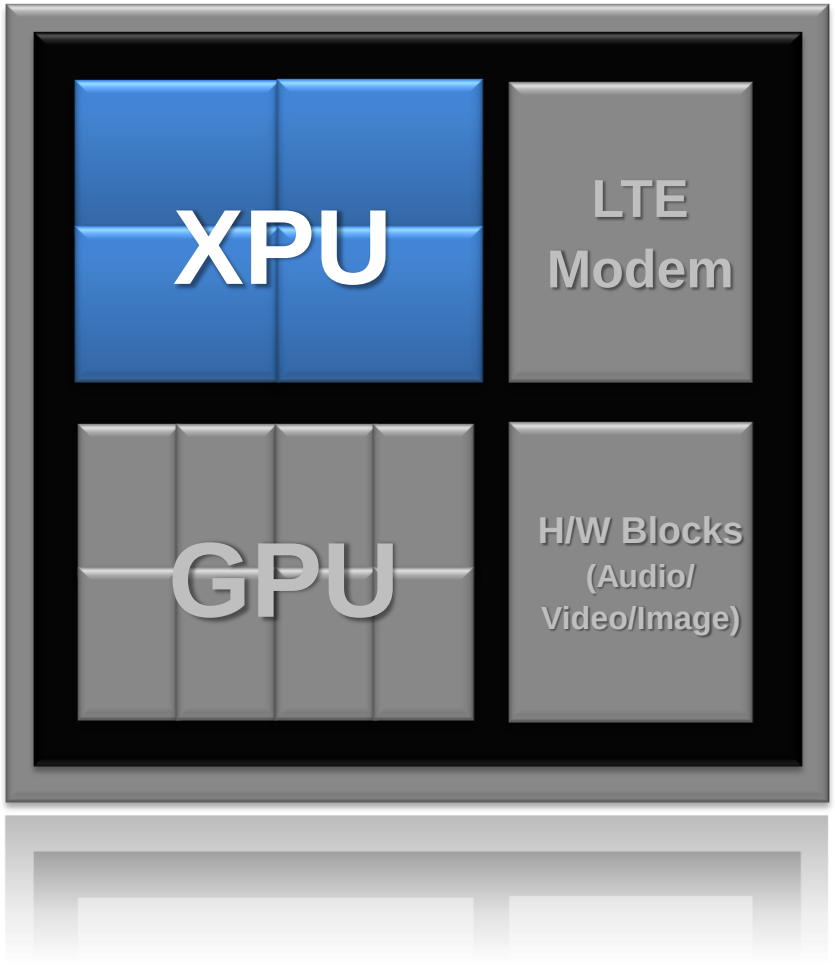
- **Inadequate FP Performance**
  - Flagship mobile GPU: 200~300 GFLOPS
  - Real-time ray tracing @HD: >300Mray/sec (1~2TFLOPS)
- **Unsuitable Execution Model**
  - “Multithreaded SIMD (SIMT)” is not fit for rendering incoherent rays
  - Tree construction is a irregular work - sorting and random memory access
- **Weak Branch Supports**
  - Performance drops when recursion, function calls, control flow…

Ray Tracing on Fully Dedicated H/W



- Performance & power-efficient
- Full H/W specialization only for ray tracing (ray generation, tree traversal, shading, and BVH construction [Michael et al SIGGRAPH 2013])
- Several architectures have been proposed for the PC environment
  - SaarCOR [Schmittler et al. 2004]
  - RPU [Woop et al. 2005]
  - …
- Low Flexibility

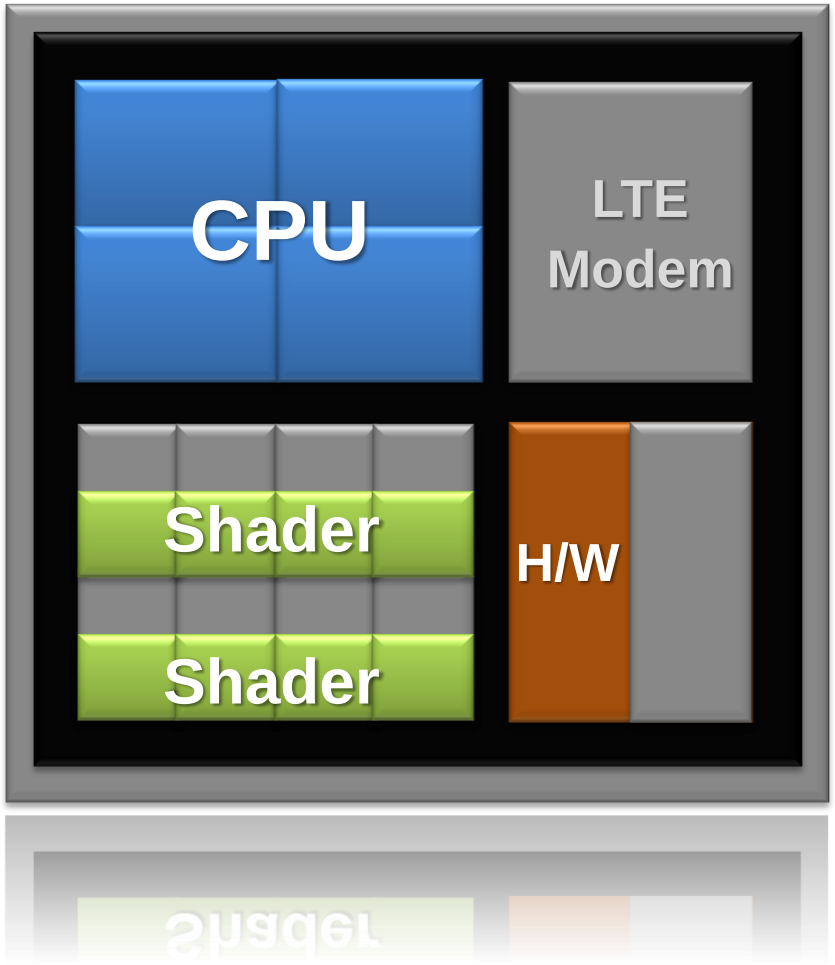
Fully S/W Ray Tracing on New Processor?



- Highly Flexible
  - support fully programmable tree construction and rendering
- Reconfigurable SIMT processor [Kim at al, 2012]
  - Configured to both SIMT/MIMD modes
- MIMD threaded multi-processor [Spjut, 2012]
  - Mobile version of the TraX MIMD T/M [Kopta et al. 2010]
- Performance / Power not enough

Proposed System Architecture

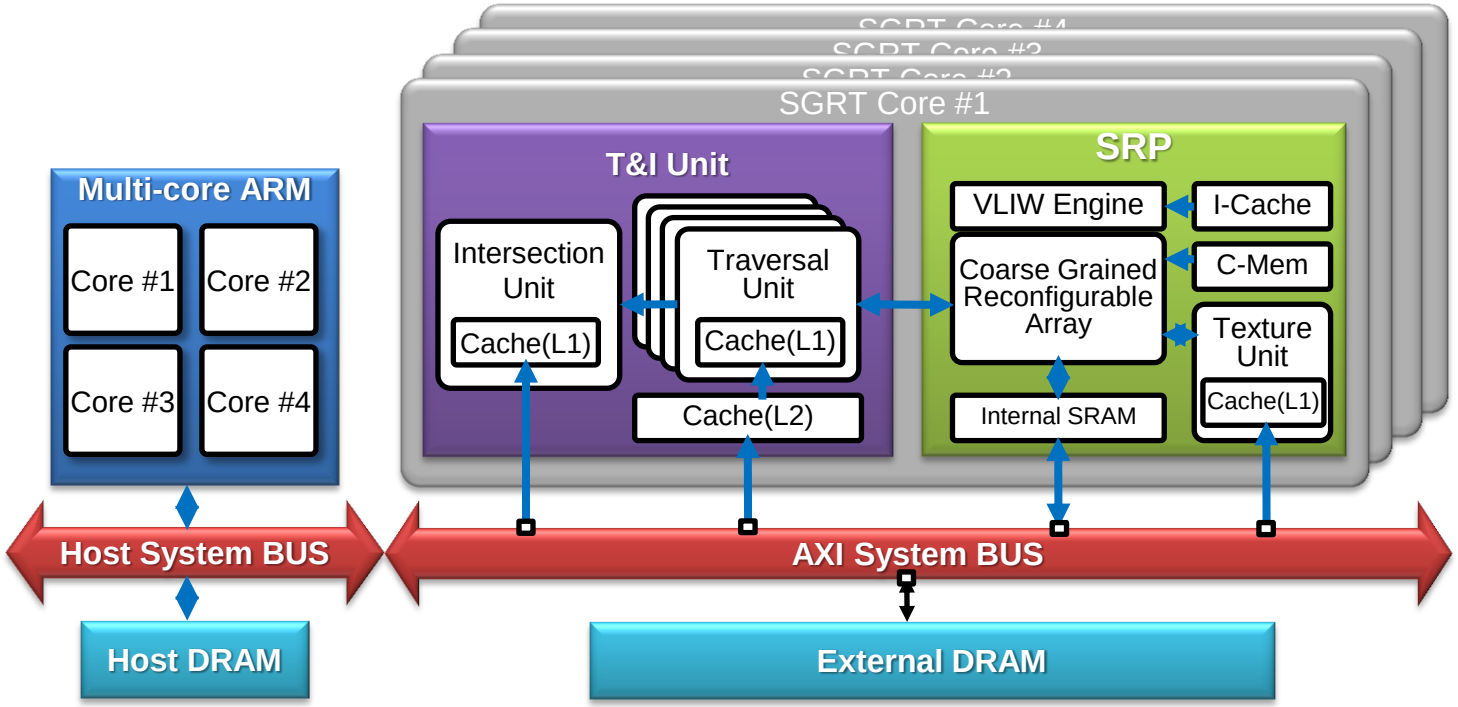
Design Decision



- **Hybrid S/W and H/W solution with existing CPUs/GPUs**
  - Tree Build: *sorting, irregular work* → Multi-core CPU
  - Traversal & Intersection (T&I): *embarrassingly parallel* → Dedicated H/W
  - Ray Gen. & Shading (RGS): *need for flexibility* → Programmable shader
- **Acceleration Structure : BVH**
- **Single Ray-based Architecture**
  - More robust for incoherent rays than SIMD packet tracing.

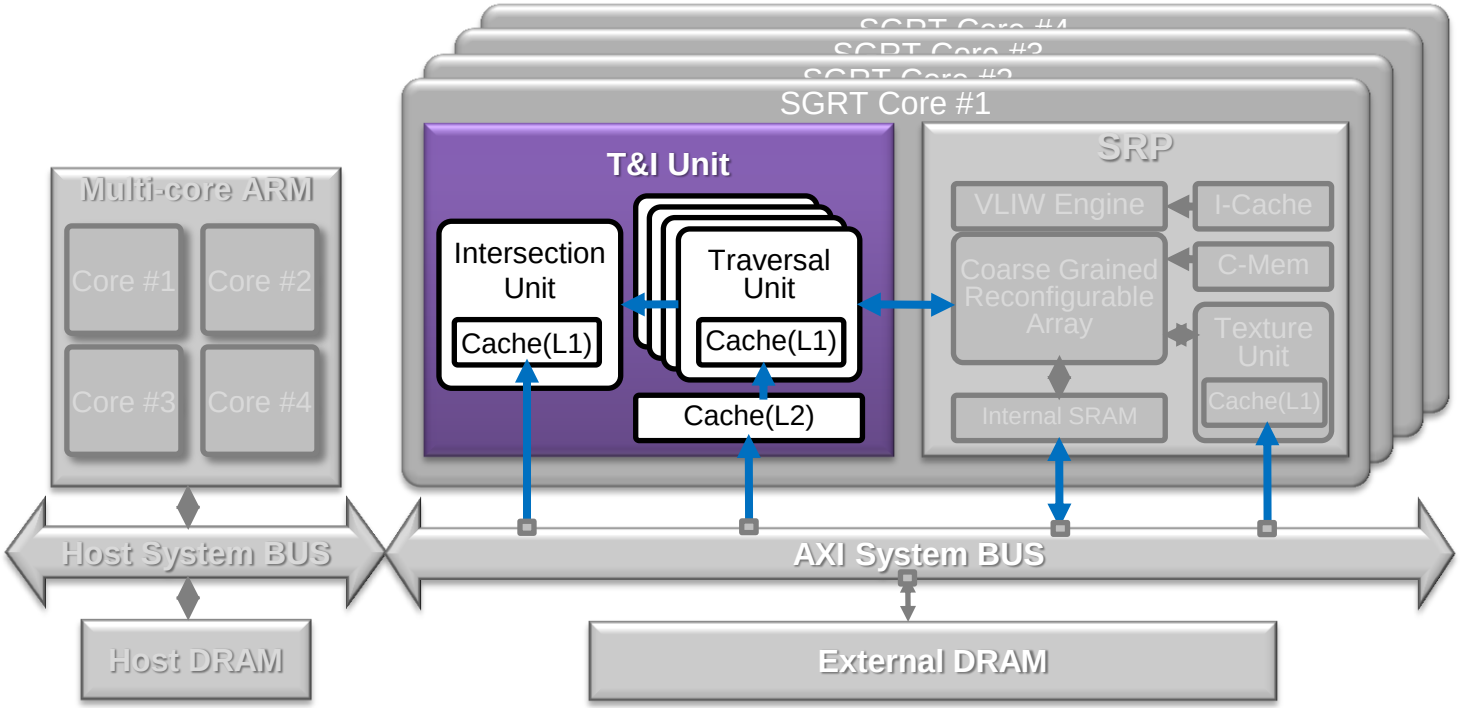
Overall System Architecture

- SGRT = T&I Unit + SRP
- T&I Unit : A fast compact hardware engine that accelerates a “Traversal and Intersection” operation, based on [Nah et al, 2011]
- SRP : A flexible reconfigurable processor that supports software “Ray Generation and Shading” [Lee et al, 2011/2012]

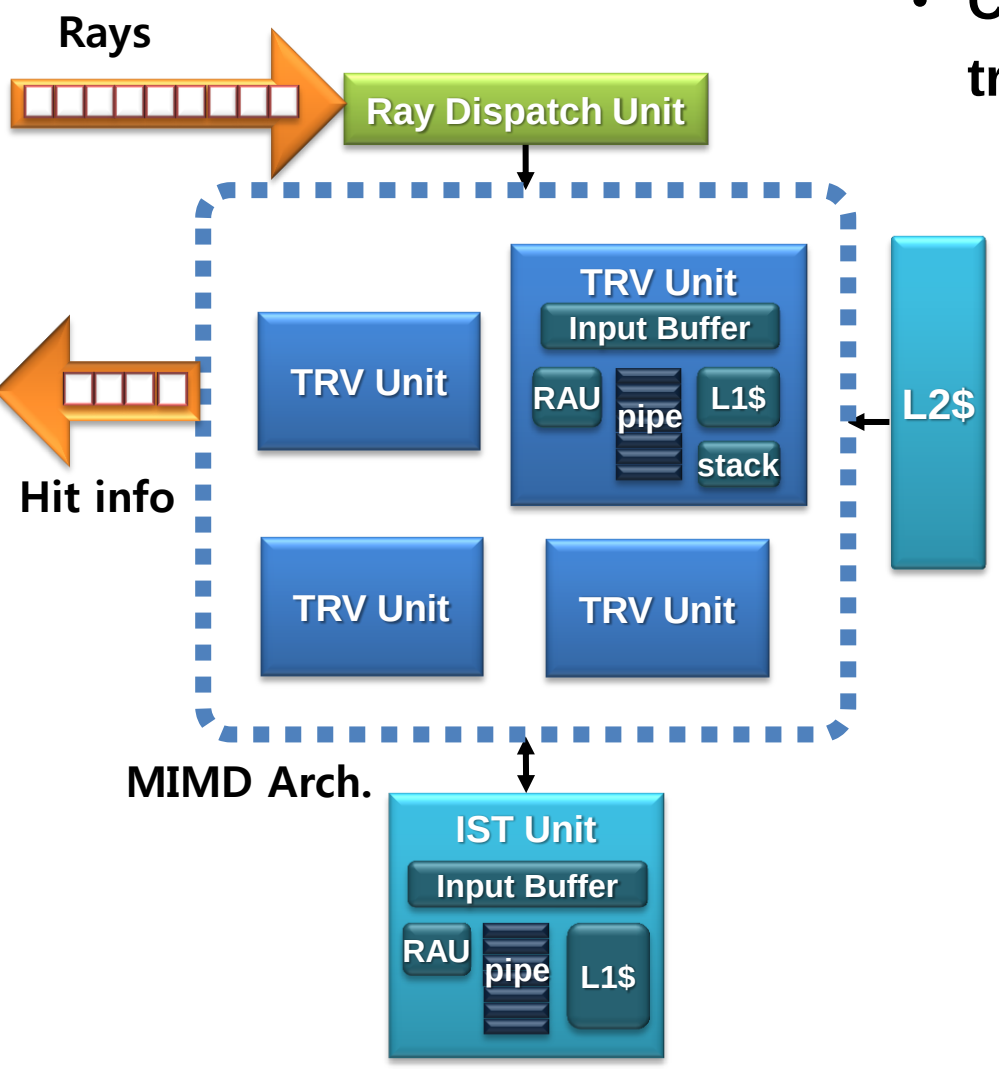


Overall System Architecture

- SGRT = T&I Unit + SRP
- T&I Unit : A fast compact hardware engine that accelerates a “Traversal and Intersection” operation
- SRP : A flexible reconfigurable processor that supports software “Ray Generation and Shading”

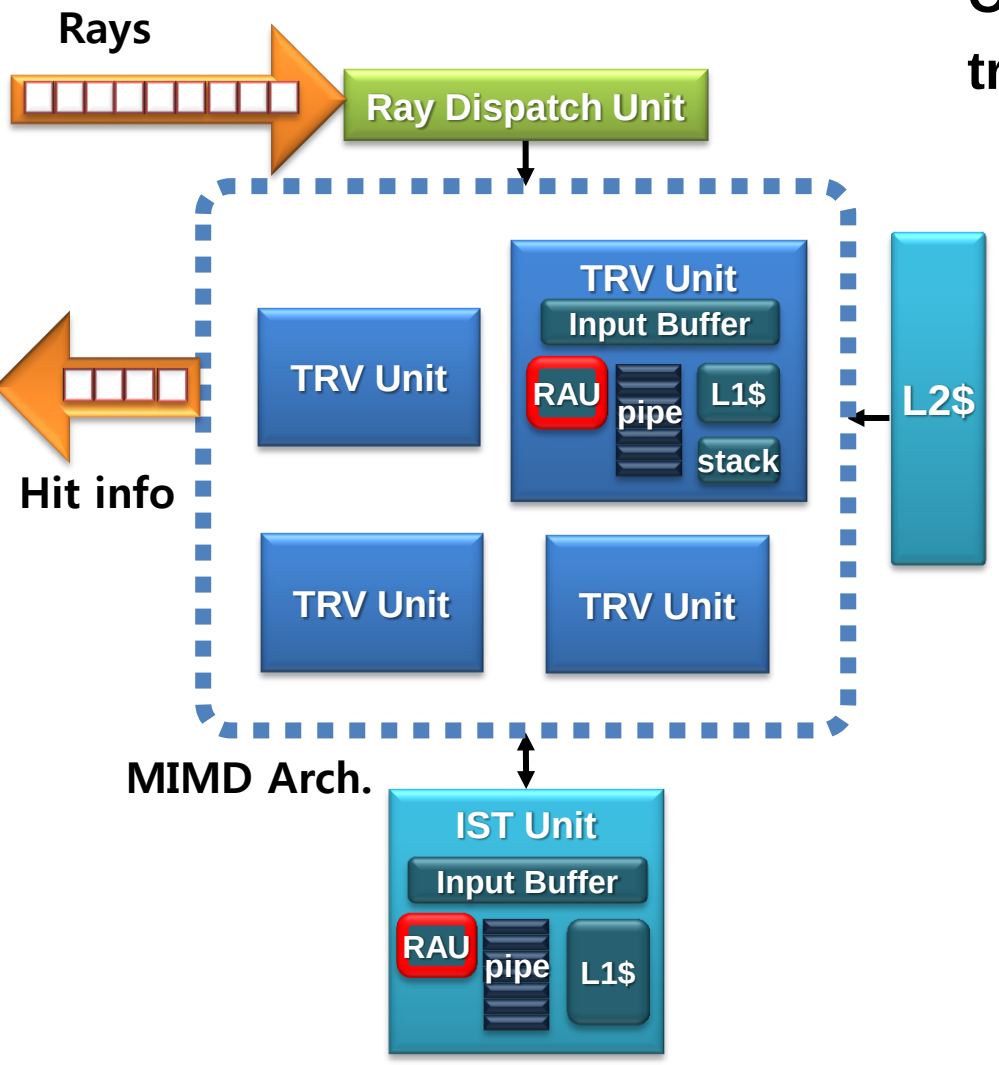


T&I Unit : A MIMD H/W Accelerator



- **Compact & fast H/W accelerator for traversal / intersection**
  - Revision of T&I Engine
    [Nah, SIGGRAPH Asia 2011]
    - Single-ray-based MIMD arch.
    - Ray Accumulation Unit (RAU)
      : H/W multithreading
    - Decoupled memory & computation pipeline
  - *Parallel Pipelined TRV Unit*
    [Kim, SIGGRAPH Asia 2012]
  - *Early Intersection Test*
    : Pre-filtering for skipping IST

T&I Unit : A MIMD H/W Accelerator

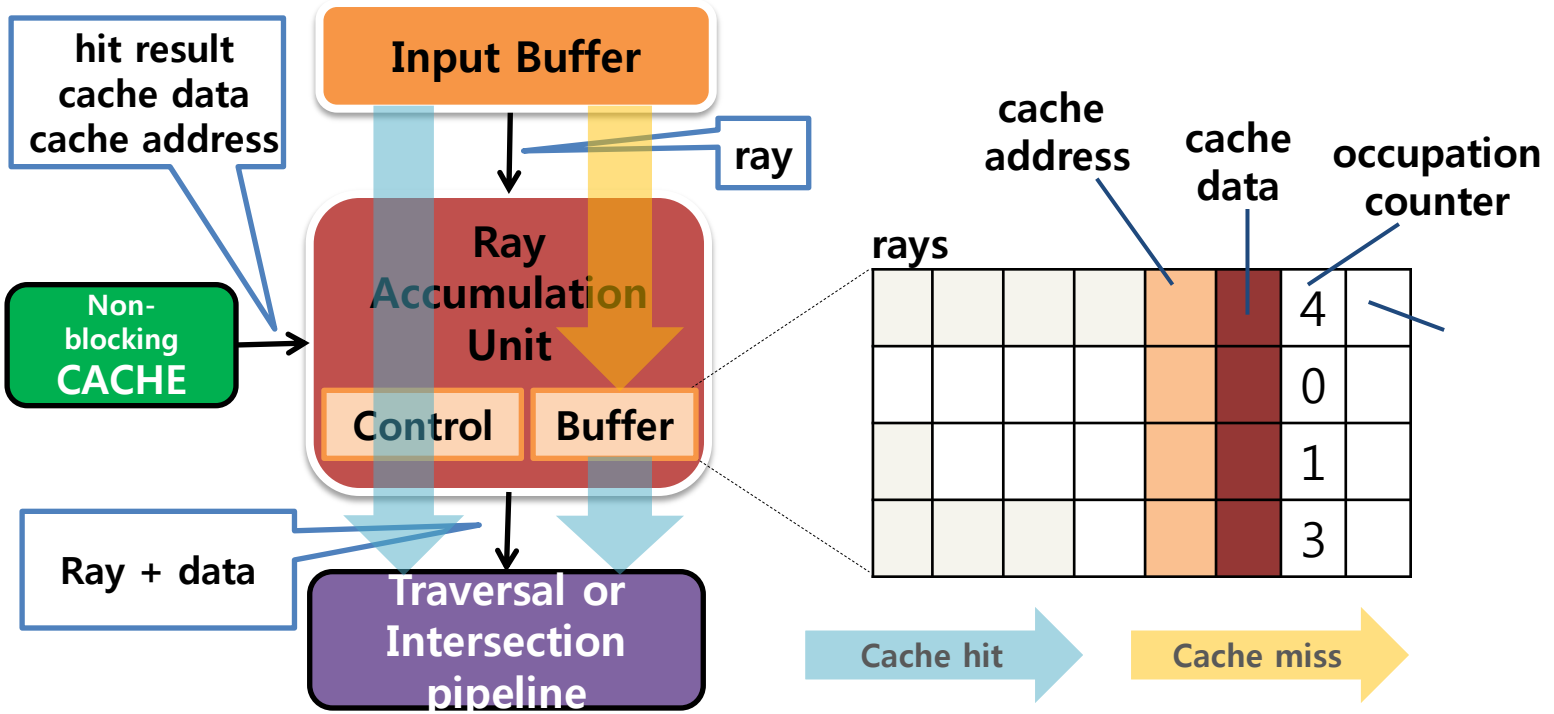


- **Compact & fast H/W accelerator for traversal / intersection**
  - Revised T&I Engine
    [Nah, SIGGRAPH Asia 2011]
  - Single-ray-based MIMD arch.
  - **Ray Accumulation Unit (RAU) : H/W multithreading**
  - Decoupled memory & computation pipeline

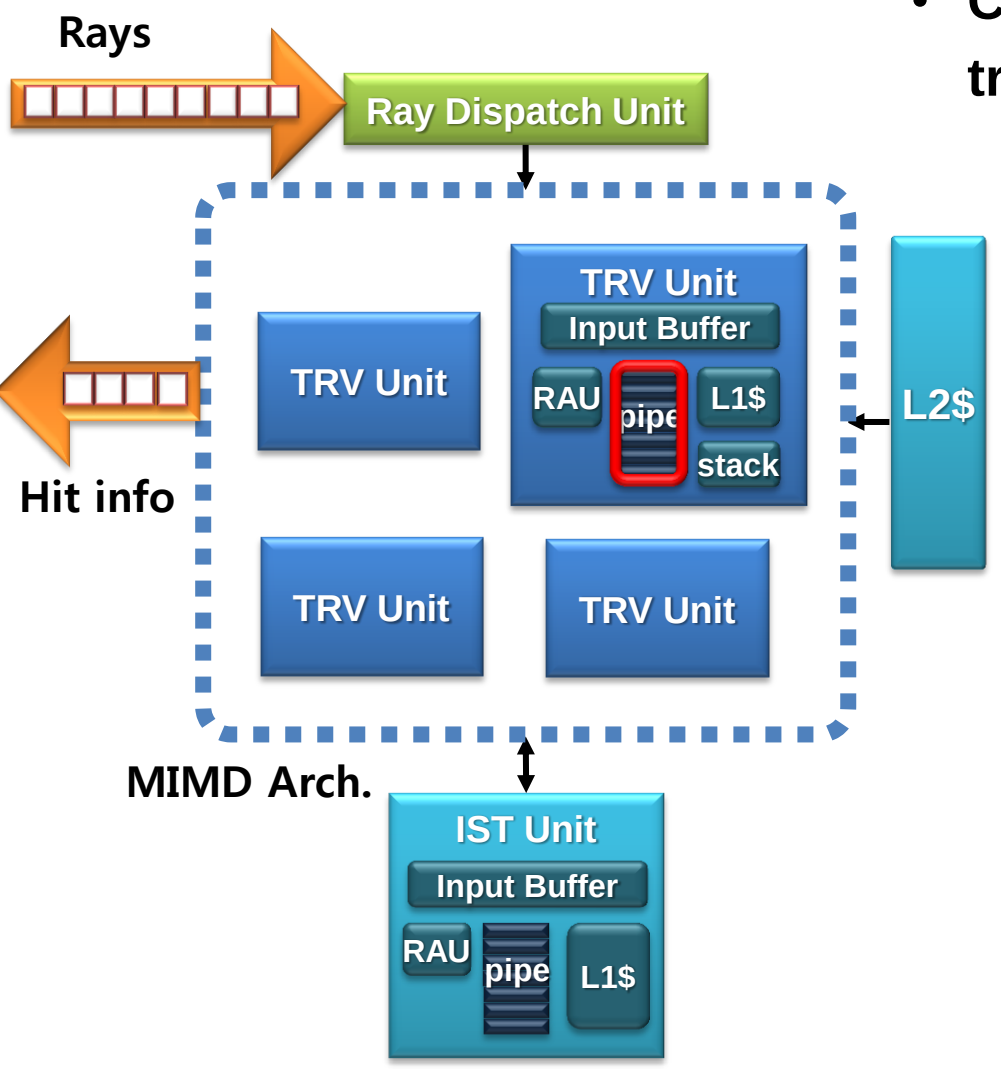
  - *Parallel Pipelined TRV Unit*
    [Kim, SIGGRAPH Asia 2012]
  - *Early Intersection Test*
    : Pre-filtering for skipping IST

Ray Accumulation Unit

- Specialized H/W multi-threading for latency hiding
  - Cache missed rays are accumulated in RA buffer, other rays can be processed during this period
  - Coherence can be increased, the rays that reference the same cache line are accumulated in the same row in an RA buffer

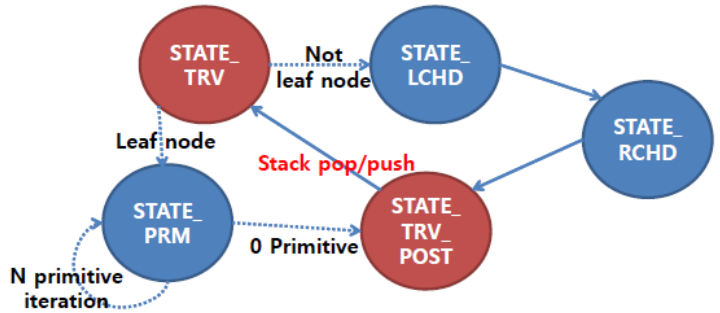


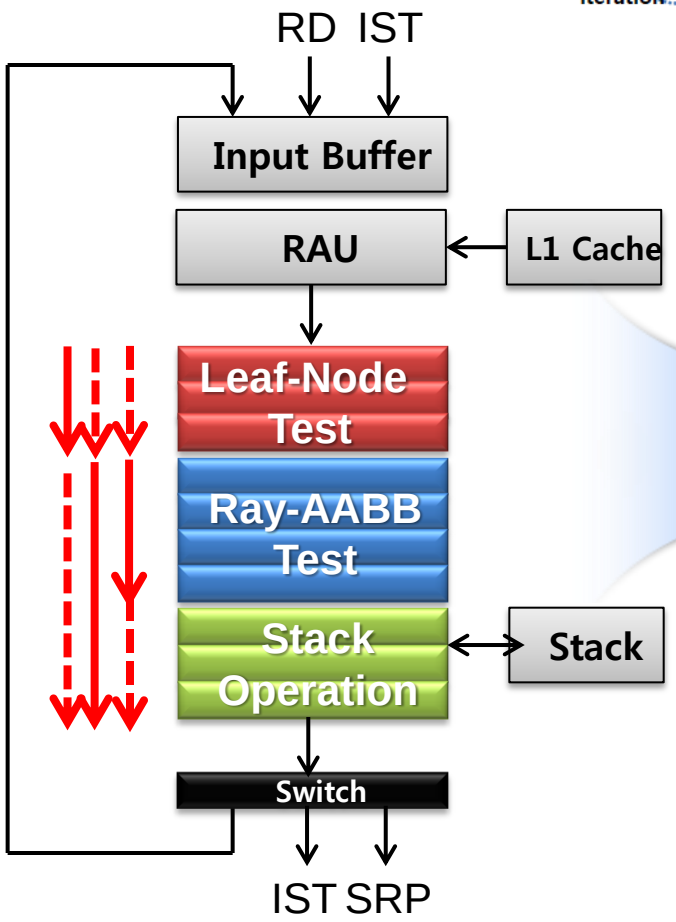
T&I Unit : A MIMD H/W Accelerator

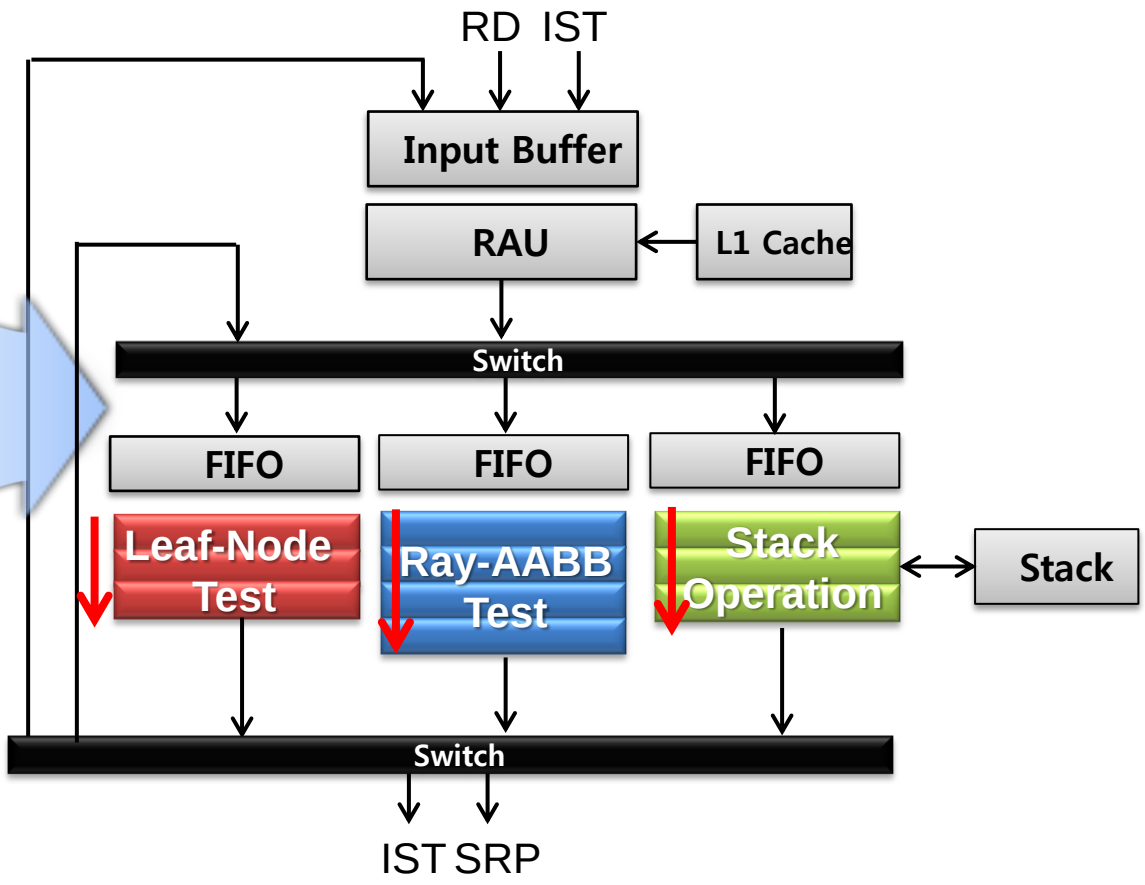


- **Compact & fast H/W accelerator for traversal / intersection**
  - Revised T&I Engine
    [Nah, SIGGRAPH Asia 2011]
    - Single-ray-based MIMD arch.
    - Ray Accumulation Unit (RAU)
      : H/W multithreading
    - Decoupled memory & computation pipeline
  - ***Parallel Pipelined TRV Unit***
    [Kim, SIGGRAPH Asia 2012]
  - *Early Intersection Test*
    : Pre-filtering for skipping IST

Parallel Pipelined TRV Unit

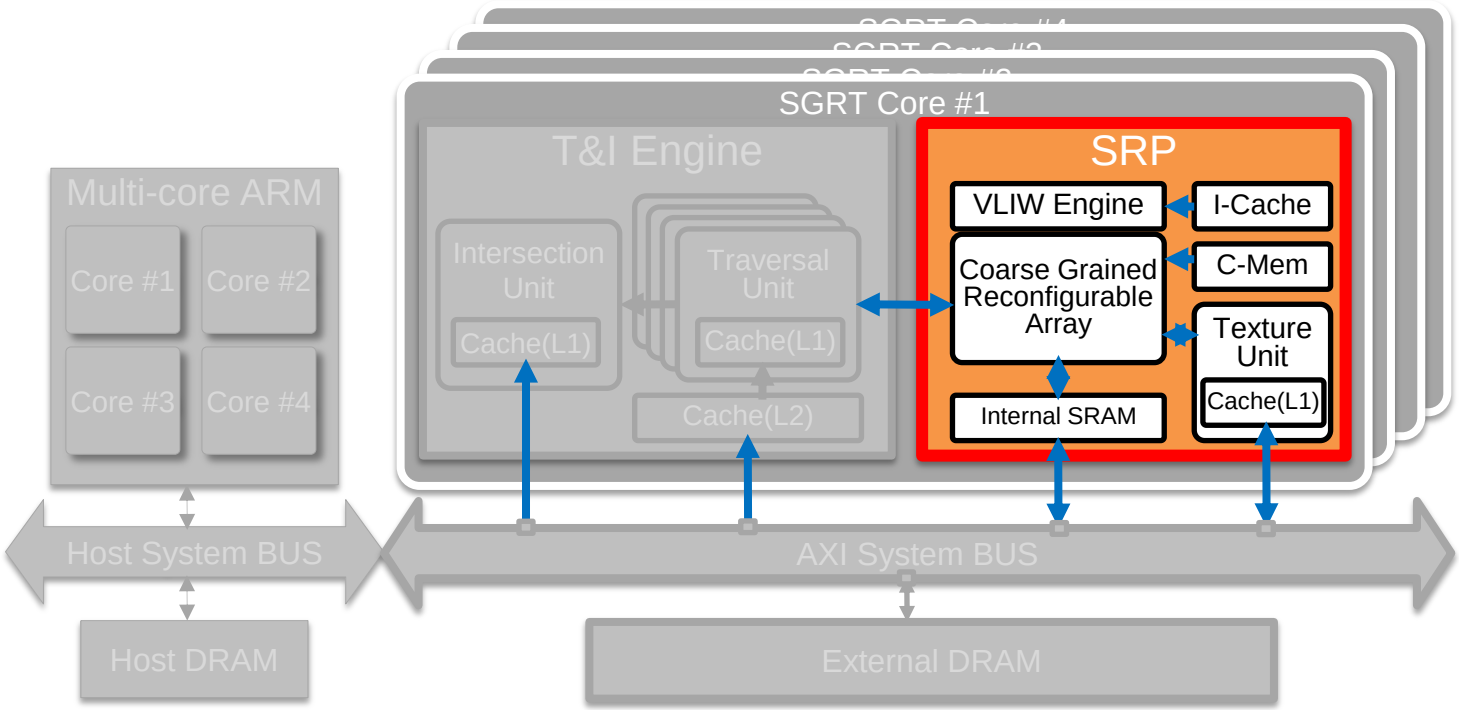






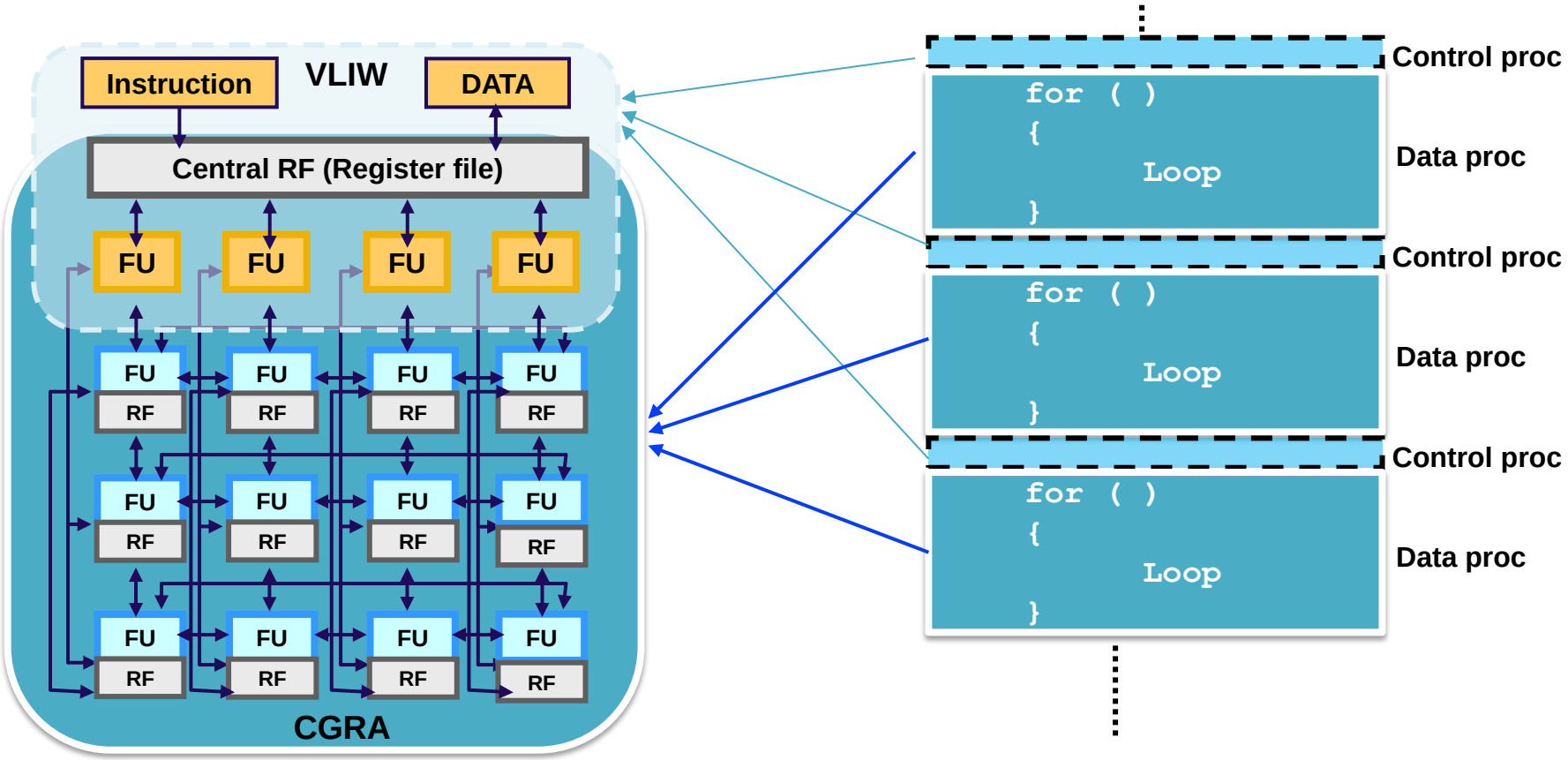
Overall System Architecture

- SGRT = T&I Unit + SRP
- T&I Unit : A fast compact hardware engine that accelerates a “Traversal and Intersection” operation
- SRP : A flexible reconfigurable processor that supports software “Ray Generation and Shading”

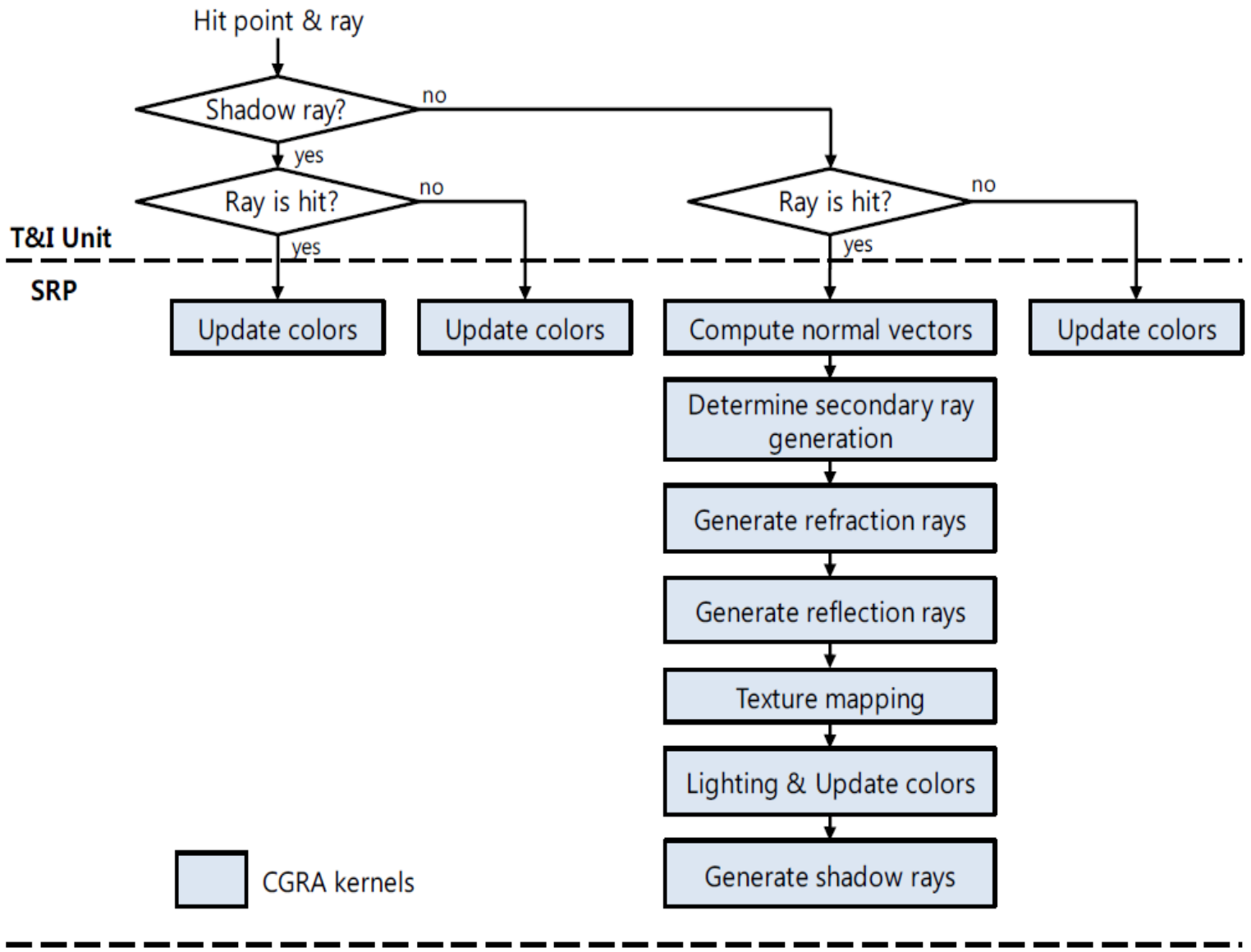


Reconfigurable Processor

- A flexible architecture template [Lee, HPG 2011/2012]
- ISA such as arithmetic, special function and texture are properly implemented.
- The VLIW engine useful for GP computations (function invocation, control flow).
- The CGRA makes full use of software pipeline technique for loop acceleration.



Execution Flow of Shading & Ray Generation



Results and Analysis
(Cycle Accurate Simulation)

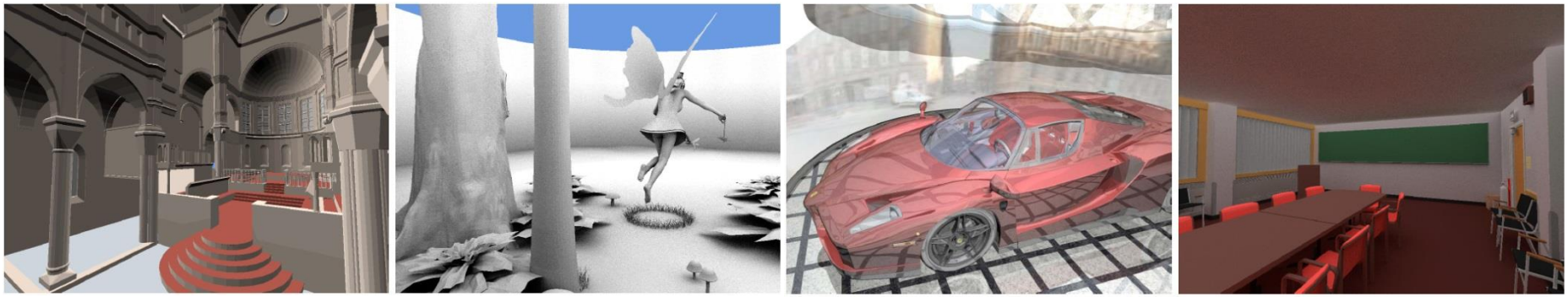
Cycle Accurate Simulation

- **Simulation environment**
  - T&I : In-house cycle-accurate simulator + GDDR memory simulator [GPGPUsim]
  - RGS Kernels : In-house compiled simulator, “CSim”

- **H/W setup**
  - SGRT uses four cores (four T&I units and four SRP cores).
  - T&I unit, the number of TRVs and ISTs = 4:1
  - Clock frequency for the T&I unit and the SRP at 500 MHz and 1 GHz
  - 1 GHz clock and 32-bit 2-channel GDDR memory (close to LPDDR3 memory)

Cycle Accurate Simulation

- Test scenes
  - Four test scenes : Sibenik (80 K triangles), Fairy (174 K triangles), Ferrari (210 K triangles), and Conference (282 K triangles).
  - "Primary ray", "Ambient occlusion ray", "Diffuse inter-reflection ray" [Aila, HPG 2012], "Forced specular ray (2-bounce)"
  - 1024x768



Test scenes: Sibenik (primal rays), Fairy (ambient occlusion rays), Ferrari (forced specular rays (2-bounce)), and Conference (diffuse inter-reflection rays).

Simulation Results : RGS

- RGS Performance
  - 147-198 Mray/sec
  - Texture cache concerns : Mip-mapping & Compression

Test scene	Ray type	Cache hit rate (%) Texture	Data	Bandwidth (GB/s)	Performance (Mrays/sec)
Sibenik	Primary	-	96.76	0.5	182.11
(80K tri.)	FSR	-	91.24	1.9	172.25
Fairy	Primary	93.25	96.87	0.8	175.66
(179K tri.)	FSR	81.49	94.91	1.9	147.45
Ferrari	Primary	86.12	98.09	0.6	183.28
(210K tri.)	FSR	75.95	95.71	2.0	163.67
Conference	Primary	-	98.44	0.2	198.32
(282K tri.)	FSR	-	95.72	0.8	158.79

Simulation Results : New TRV

- SPTRV vs. PPTRV
  - PPTRV outperforms up to 40% (26% on average)

Test scene	Ray type	Average steps SPTRV	Average steps PPTRV	Mrays/sec SPTRV	Mrays/sec PPTRV	Ratio to SPTRV
Sibenik (80K tri.)	Primary	61.30	23.12	27	33	1.24
	AO	36.55	14.87	48	56	1.15
	Diffuse	81.62	29.93	11	15	1.40
Fairy (179K tri.)	Primary	70.86	28.02	22	28	1.27
	AO	31.53	12.43	52	62	1.20
	Diffuse	51.72	18.99	19	24	1.31
Ferrari (210K tri.)	Primary	68.86	25.52	23	29	1.26
	AO	30.64	11.32	54	64	1.18
	Diffuse	92.24	59.20	20	25	1.25
Conference (282K tri.)	Primary	44.66	15.54	36	46	1.30
	AO	17.23	5.88	99	121	1.23
	Diffuse	43.06	14.59	33	44	1.35

Simulation Results : T&I

- Obtained a performance of 61~485 Mrays/s.
- T&I unit can compete with the ray tracer on the previous desktop GPU, Tesla, and Fermi architecture [Aila et al. 2012]
- The performance gap between the primary ray and the diffuse ray was narrow (except for Sibenik) because of MIMD with an appropriately sized cache architecture

Table 5: *Simulation results of T&I unit (Four units at 500 MHz clock).*

Test scene	Ray type	Utilization (%)		Average steps		Cache hit rate (%)			Bandwidth (GB/s)	Performance (Mrays/sec)	Ratio to Tesla	Ratio to Fermi
		TRV	IST	TRV	IST	TRV L1	TRV L2	IST L1				
Sibenik (80K tri.)	Primary	89	52	23.12	3.10	99	41	99	1.1	132	1.13	0.54
	AO	92	55	14.87	1.17	99	68	99	0.1	222	1.86	0.91
	Diffuse	49	38	29.93	4.50	72	65	88	2.6	61	1.30	0.65
Fairy (179K tri.)	Primary	67	67	28.02	4.40	97	49	99	1.5	112	1.50	0.73
	AO	61	79	12.43	1.98	94	84	99	0.4	249	2.69	1.52
	Diffuse	49	67	18.99	4.21	73	63	91	3.7	97	2.38	1.33
Conference (282K tri.)	Primary	78	70	15.54	3.35	99	57	99	0.3	184	1.30	0.68
	AO	86	55	11.76	0.12	99	63	99	0.1	485	3.61	1.71
	Diffuse	62	64	14.59	3.82	90	70	96	3.3	178	2.92	1.41

Simulation Results : Overall

- 4-SGRT cores includes all the T&I units and the SRPs
- Better performance compared to PC and mobile solution in terms of perf / area

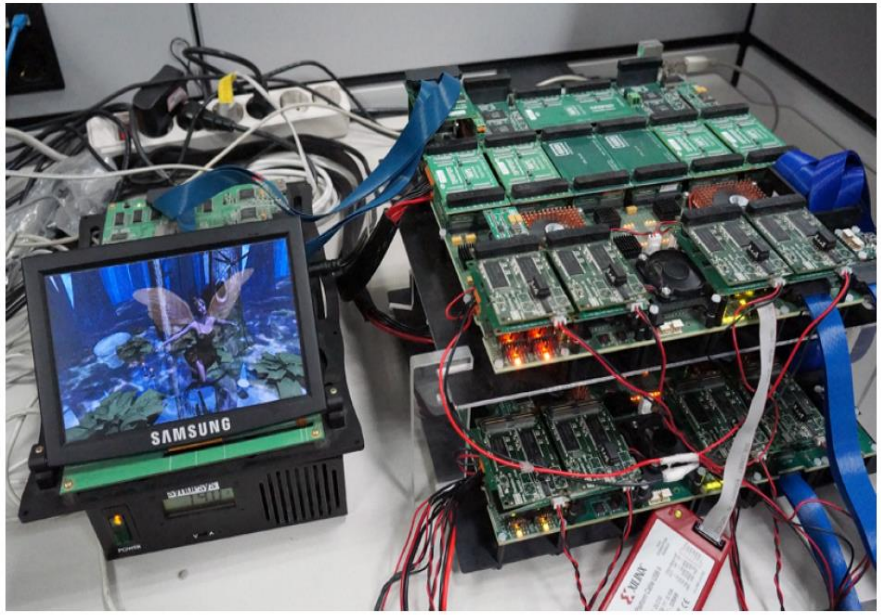
Table 6: *Performance comparison for the Fairy scene.*

	Desktop		Mobile		
	GPU (Optix) [OptiX 2013]	MIC [Wald 2012]	Multiprocessor [Spjut et al. 2012]	Multiprocessor [Kim et al. 2012a]	Ours
Parallelism	SIMT	Wide SIMD	MIMD	SIMT/MIMD	MIMD + loop parallelism
Resolution	1920×1200	1920×1200	1280×720	1280×720	1920×1080 (Full HD)
Platform	NVIDIA GeForce GTX 680	Intel MIC	Thead Multiprocessor	Reconfigurable SIMT	SGRT (H/W + shader)
Clock (MHz)	1006	1000	500	400	500 (H/W), 1000 (shader)
Process (nm)	28	45	65	90	65
Area (mm^2)	294	-	25	16	25
BVH type	SAH	SAH	BVH	-	SAH
FPS	37	29	9	2	34

Results and Analysis
(FPGA Prototyping)

FPGA Prototypes : Validation

- Implemented with Verilog/RTL codes
- FPGA prototype
  - Xilinx Virtex-6 LX760 FPGA chips, Synopsys HAPS-64 board.
  - In-house fabricated LCD display (800x480) board.
- A single SGRT core is partitioned and mapped to two Virtex-6 chips
  - Due to the size limitation of an FPGA chip.
  - including a T&I unit, SRP core, AXI-bus, and memory controller
  - Synthesized and implemented with Xilinx Tools at a 45 MHz clock frequency.



Conclusion

Conclusion

- **SGRT : Mobile Ray Tracing GPU**
  - T&I unit + SRP
  - Key features : MIMD arch, RAU, Parallel TRV, Early IST
  - Evaluated through a cycle-accurate simulation and verified by FPGA prototyping.

- **Future Work**
  - Support complex shading with advanced shader cores,
  - Programmable IST units
  - Low-power architecture
  - ASIC migration