Squeezing Every Drop Of Performance Out Of The iPhone

Noel Llopis Snappy Touch

iPhoneGames

http://twitter.com/snappytouch <u>noel@snappytouch.com</u> http://gamesfromwithin.com



In the last 11 years I've made games for almost every major platform out there







РСАЧЕТАТІОП Э







In the last 11 years I've made games for almost every major platform out there



snappy touch

During that time, working on engine technology and trying to achieve maximum performance Performance always very important





Almost a year ago I started Snappy Touch





Almost a year ago I started Snappy Touch



Running at 40 FPS on first attempt...



Running at 40 FPS on first attempt...

... on the simulator!







Lessons learned getting frame rate from 4 fps to ~30 fps

• Performance analysis tools



- Performance analysis tools
- CPU



- Performance analysis tools
- CPU
- Game loop architecture



- Performance analysis tools
- CPU
- Game loop architecture
- Rendering



- Performance analysis tools
- CPU
- Game loop architecture
- Rendering
- Memory



Performance Analysis Tools





• Turning features on and off in real time



- Turning features on and off in real time
- Works very well to get very rough estimates



- Turning features on and off in real time
- Works very well to get very rough estimates
- Simulation vs. geometry creation vs. rendering

















- Versatile
- Can display frame rate, object allocations, memory usage, or even OpenGL calls.



- Versatile
- Can display frame rate, object allocations, memory usage, or even OpenGL calls.
- Rut it on the real device, not the simulator!





Best way to start.

Frame rate over time plus CPU sampler



Digging deeper



Digging deeper



Also good for tracking memory leaks (wish for more automated way though)



Also good for tracking memory leaks (wish for more automated way though)



• First pass with Instruments confirmed that Flower Garden was simulation bound.



- First pass with Instruments confirmed that Flower Garden was simulation bound.
- Especially the matrix multiplies and geometry creation sections.








- CPU usage
- More in depth than Instruments



- CPU usage
- More in depth than Instruments
- Can even report good info on cache misses



1	Self	Total 🔻	Library Name	Symbol Name	
	1.4%	1.4%	Sketch	-[SKTGraphicView drawRect:]	A
	1.3%	1.3%	Sketch	-[SKTGraphic drawInView:isSelected:]	0
	1.0%	1.0%	AppKit	▶ dyld_stub_objc_msgSend	
	0.8%	0.8%	AppKit	► +[NSBezierPath clipRect:]	
	0.7%	0.7%	AppKit	► -[NSBezierPath stroke]	
	0.7%	0.7%	commpage	▶pthread_getspecific	
	0.6%	0.6%	Sketch	-[SKTGraphic drawHandleAtPoint:inView:]	
	0.6%	0.6%	commpage	▶bzero	
Į.	0 64	U 24	Annkit	▶_INCRaziarDath/NCRaziarDathDavicaDrimitivae\ addDathC	
1	Self	Total 🔻	Library Name	4%) process samples selected Symbol Name	_
	0.0%	99.6%	dvld	▼_dyld_start	A
	0.0%	99.6%	Sketch	▼_start	0
	0.0%	99.6%	Sketch	▼ main	•
	0.0%	99.6%	AppKit	▼ NSApplicationMain	
	0.0%	99.6%	AppKit	▼-[NSApplication run]	
	0.0%	99.6%	AppKit	▼-[NSApplication sendEvent:]	
	0.0%	99.6%	AppKit	▼-[NSWindow sendEvent:]	
	0.0%	99.6%	AppKit		<u> </u>
1	n n42	99 69	Annkit	w_[NSTitladErama_mouraDown-]	1
			4511 of 4511 (10	0.0%) process samples displayed	
		katch [403]	The The		





CPU Debug Optional VFP Coprocessor Interface ARM Instruction Cache TrustZone™ enabled TCRAM 0 ARM11[™] core TCRAM 1 S 1 LL., N Memory Management 9 AMBA AXI Interface --Instruction Data

Interface

Interface

Σ

8 4

Controller

Data

Cache

TCRAM 0

TCRAM 1

Peripheral

Port

snappy touch

DMA

This is where the biggest bottlenecks were for Flower Garden at first

CPU

• 32-bit RISC ARM 11



This is where the biggest bottlenecks were for Flower Garden at first

• 32-bit RISC ARM 11

• 400-535Mhz



CPU

- 32-bit RISC ARM 11
- 400-535Mhz
- iPhone 2G/3G and iPod Touch 1st and 2nd gen



CPU (iPhone 3GS)

S5PC100 Block Diagram





CPU (iPhone 3GS)

• Cortex-A8 600MHz

S5PC100 Block Diagram





CPU (iPhone 3GS)

- Cortex-A8 600MHz
- More advanced architecture

S5PC100 Block Diagram











• CPU has a special thumb mode.





- CPU has a special thumb mode.
- Less memory, maybe better performance.







- CPU has a special thumb mode.
- Less memory, maybe better performance.
- No floating point support.





- CPU has a special thumb mode.
- Less memory, maybe better performance.
- No floating point support.
- It's on by default!





- CPU has a special thumb mode.
- Less memory, maybe better performance.
- No floating point support.
- It's on by default!
- Potentially HUGE wins turning it off.





- CPU has a special thumb mode.
- Less memory, maybe better performance.
- No floating point support.
- It's on by default!
- Potentially HUGE wins turning it off.

Show Warnings		
ser-Defined		
GCC_C_LANGUAGE_STANDARD	c99	
GCC_DYNAMIC_NO_PIC	NO	
GCC_OBJC_CALL_CXX_CDTORS	YES	h
GCC_OPTIMIZATION_LEVEL	0	I
CCC PRECOMPLIE PREEIX HEADER	VES	н
dee_rkeeowniee_rkerix_neAbek	125	
CCC_PREFIX_HEADEK	FlowerGarden_Prefix.pch	I
CCC_PREFIX_HEADER GCC_THUMB_SUPPORT	FlowerGarden_Prefix.pch NO	l
GCC_PREFIX_HEADER GCC_THUMB_SUPPORT GCC_WARN_ABOUT_MISSING_NEWLINE	FlowerGarden_Prefix.pch NO YES	
GCC_PREFIX_HEADER GCC_THUMB_SUPPORT GCC_WARN_ABOUT_MISSING_NEWLINE GCC_WARN_ABOUT_RETURN_TYPE	FlowerGarden_Prefix.pch NO YES YES	
GCC_THUMB_SUPPORT GCC_WARN_ABOUT_MISSING_NEWLINE GCC_WARN_ABOUT_RETURN_TYPE GCC_WARN_CHECK_SWITCH_STATEMENTS	FlowerGarden_Prefix.pch NO YES YES YES	





• Turning off Thumb mode increased performance in Flower Garden by over 2x



- Turning off Thumb mode increased performance in Flower Garden by over 2x
- Heavy usage of floating point operations though



- Turning off Thumb mode increased performance in Flower Garden by over 2x
- Heavy usage of floating point operations though
- Most games will probably benefit from turning it off (especially 3D games)



Integer Divide



So plan accordingly!

Integer Divide



There is no integer divide





• The main CPU has no floating point support.



- The main CPU has no floating point support.
- Compiled C/C++/ObjC code uses the vector floating point unit for any floating point operations.



- The main CPU has no floating point support.
- Compiled C/C++/ObjC code uses the vector floating point unit for any floating point operations.
- Can program the VFP in assembly for max performance.





• Big win when many floating point operations can be vectorized.



- Big win when many floating point operations can be vectorized.
- Matrix multiplies, skinning, etc



- Big win when many floating point operations can be vectorized.
- Matrix multiplies, skinning, etc
- vfpmath project in Google Code



```
void Matrix4Mul(const float* src_mat_1, const float* src_mat_2,
float* dst_mat)
{
    asm volatile (VFP_SWITCH_TO_ARM
        VFP_VECTOR_LENGTH(3)
        "fldmias %2, {s8-s23} \n\t"
        "fldmias %1!, {s0-s3} \n\t"
        "fmuls s24, s8, s0 \n\t"
        "fmacs s24, s12, s1 \n\t"
//...
```


Game Loop Architecture







Main difference between games and apps is the constant amount of "stuff" happening on screen.



Gather input





Main difference between games and apps is the constant amount of "stuff" happening on screen.







Main difference between games and apps is the constant amount of "stuff" happening on screen.



Main difference between games and apps is the constant amount of "stuff" happening on screen.



Main difference between games and apps is the constant amount of "stuff" happening on screen.

Event-driven architecture is not a very good match for games

OpenGL



OK for games that don't require a high/steady frame rate

• Easiest way to drive the main loop



OK for games that don't require a high/steady frame rate

- Easiest way to drive the main loop
- All the samples from Apple



- Easiest way to drive the main loop
- All the samples from Apple

```
m_gameLoopTimer = [NSTimer
scheduledTimerWithTimeInterval:animationInterval
target:self
selector:@selector(doFrame)
userInfo:nil
repeats:YES];
```





















The app becomes even less responsive















- Call it at a higher frequency so main loop is called right away.
- Unfortunately that floods the message queue and causes delay on touch events





• Put the game on a separate thread from the UI one.



- Put the game on a separate thread from the UI one.
- Runs as fast as possible without delays





Careful what you do when parsing input events from main thread



- Careful what you do when parsing input events from main thread
- Can be more complex to debug



- Careful what you do when parsing input events from main thread
- Can be more complex to debug
- Running as fast as possible might not always be desirable (draining battery life)





 Can also split game update and rendering in two different threads.


Threads

- Can also split game update and rendering in two different threads.
- Syncing the two is much more difficult



Threads

- Can also split game update and rendering in two different threads.
- Syncing the two is much more difficult
- Can only use one thread for each OpenGL context



Threads

- Can also split game update and rendering in two different threads.
- Syncing the two is much more difficult
- Can only use one thread for each OpenGL context
- Theoretically best results, but maybe not much difference



Thread Driving Loop



That's what I'm using in Flower Garden

Thread Driving Loop



Thread Driving Loop

```
- (void)mainLoopTimer
{
    while (![[NSThread currentThread] isCancelled])
    {
        [self performSelectorOnMainThread:@selector(doFrame)
            withObject:nil waitUntilDone:YES];
        usleep(2000);
    }
}
```





• Can be best of both worlds



- Can be best of both worlds
- Consistent call to main loop



- Can be best of both worlds
- Consistent call to main loop
- No flooding of events (cuts in line)





• Triggered by display refresh



- Triggered by display refresh
- Can choose to get a call every X updates



- Triggered by display refresh
- Can choose to get a call every X updates
- Consistent update



- Triggered by display refresh
- Can choose to get a call every X updates
- Consistent update
- Perfect for games



- Triggered by display refresh
- Can choose to get a call every X updates
- Consistent update
- Perfect for games

m_displayLink = [CADisplayLink displayLinkWithTarget: selfselector:@selector(mainLoop:)]; m_displayLink.frameInterval = 2; [m_displayLink addToRunLoop :[NSRunLoopcurrentRunLoop] forMode:NSDefaultRunLoopMode];





• Only available on SDK 3.1



- Only available on SDK 3.1
- Definitely the way to go for future games









iPhone 2G iPhone 3G iPod Touch 1st, 2nd, and some 3rd gen







































snappy touch











Avoid locking some resource while the GPU is rendering.







- Avoid locking some resource while the GPU is rendering.
- Might get a small speed up by first presenting the frame, then updating game state, and finally rendering.







- Avoid locking some resource while the GPU is rendering.
- Might get a small speed up by first presenting the frame, then updating game state, and finally rendering.
- Almost no difference in Flower Garden





• A lot of the usual advice for GPUs:



- A lot of the usual advice for GPUs:
- Minimize draw primitive calls



- A lot of the usual advice for GPUs:
- Minimize draw primitive calls
- Minimize state changes


Rendering

- A lot of the usual advice for GPUs:
- Minimize draw primitive calls
- Minimize state changes





 Rendering vertices can become a bottleneck



- Rendering vertices can become a bottleneck
- No true Vertex Buffer Objects on the MBX, so CPU involved in every draw call



- Rendering vertices can become a bottleneck
- No true Vertex Buffer Objects on the MBX, so CPU involved in every draw call
- Fixed on the 3GS though





• Make vertices as small as possible



• Make vertices as small as possible

```
struct PetalVertex
{
    #ifdef PETAL_SMALL_VERTEX
        int16_t x, y, z;
        int16_t u, v;
        int16_t u1, v1;
#else
        float x, y, z;
        float u, v;
        float u1, v1;
#endif
};
```



• Make vertices as small as possible

```
struct PetalVertex
{
    #ifdef PETAL_SMALL_VERTEX
        int16_t x, y, z;
        int16_t u, v;
        int16_t u1, v1;
#else
        float x, y, z;
        float u, v;
        float u1, v1;
#endif
};
```





• Align vertices on at least 4-byte boundaries



- Align vertices on at least 4-byte boundaries
- Experiment with larger alignments





You get the best performance without the degenerate verts of a strip

• Use indexed lists...



You get the best performance without the degenerate verts of a strip

- Use indexed lists...
- ... ordered as if they were strips.





You can mix and match OpenGL and UIKit



- You can mix and match OpenGL and UIKit
- But you have to be very careful



- You can mix and match OpenGL and UIKit
- But you have to be very careful
- Minimize UIKit elements on top of EAGLView



- You can mix and match OpenGL and UIKit
- But you have to be very careful
- Minimize UIKit elements on top of EAGLView
- Keep them opaque



- You can mix and match OpenGL and UIKit
- But you have to be very careful
- Minimize UIKit elements on top of EAGLView
- Keep them opaque
- OK to have smaller EAGLView



Use iPhone Features



Use iPhone Features





Use multitexturing (2 texture combiners) Point sprites

Use iPhone Features



Use multitexturing (2 texture combiners) Point sprites

























• No memory guarantees



- No memory guarantees
- No idea of how much memory will be available



- No memory guarantees
- No idea of how much memory will be available
- Not even how much memory at startup!



- No memory guarantees
- No idea of how much memory will be available
- Not even how much memory at startup!
- Sometimes I've seen the game start with as low as 3MB free memory!!!



Memory



Memory

System notifies apps when memory is running low


- System notifies apps when memory is running low
- Apps are supposed to free as much memory as they can



- System notifies apps when memory is running low
- Apps are supposed to free as much memory as they can
- Horrible situation for games!













































If you wait too long and nobody does anything, process is killed (no warning) It's a giant game of chicken!



If you wait too long and nobody does anything, process is killed (no warning) It's a giant game of chicken!



If you wait too long and nobody does anything, process is killed (no warning) It's a giant game of chicken!



• Do you prefer to allocate your memory all at once at initialization time?



- Do you prefer to allocate your memory all at once at initialization time?
- Allocate memory slowly.



- Do you prefer to allocate your memory all at once at initialization time?
- Allocate memory slowly.
- When you get a warning, wait a little.



- Do you prefer to allocate your memory all at once at initialization time?
- Allocate memory slowly.
- When you get a warning, wait a little.
- If free memory doesn't increase, release some.



- Do you prefer to allocate your memory all at once at initialization time?
- Allocate memory slowly.
- When you get a warning, wait a little.
- If free memory doesn't increase, release some.
- Repeat until enough memory



- Do you prefer to allocate your memory all at once at initia
- Allocate memory
- When you get
- If free memory some.
- Repeat until enougn memory



ease

Thank you!

Twitter: @snappytouch Email: <u>noel@snappytouch.com</u> Blog: http://gamesfromwithin.com

