AN EVOLUTION OF MOBILE GRAPHICS, V2

Michael C. Shebanow
Vice President, Advanced Processor Lab
Samsung Electronics
DISCLAIMER

- The views herein are my own
- They do not represent Samsung’s vision nor product plans
• Computing History
• The Mobile Market
• Review of GPU Tech
• GPU Efficiency
• User Experience
• Tech Challenges
• Summary
The past and the trajectory into the future…

ON THE HORIZON
WHERE HAVE WE BEEN?

1950

3/31/1951
UNIVAC I Delivered
(vacuum tubes, mag tape)
(1.9 Kops/sec)

1956

6/14/1956
TX-0 Operational
(3.6K Transistors,
256x256 memory)

1960

4/7/1964
IBM 360 Announced

1964

1965

PDP8 Announced
(~800 Kops/sec)

1965

CDC 6600
(~3 Mops/sec)

1970

1976

Cray 1 Delivered
(MECL, 166 MFLOPs)

1978

1/1/1991
Initial 2G (GSM) Cell Phones

1980

Jan 1984
Apple Mac announced

1985

May 1982
Sun-1 Launched
(M68000)

1986

1976

Cray 1 Delivered
(MECL, 166 MFLOPs)

1980

Jan 1984
Apple Mac announced

1990

7/12/1999
RIM 870 Intro

1991

6/1/1991
Initial 2G (GSM) Cell Phones

1995

6/12/2000
AMD Athlon Thunderbird
(1 GHz X86)

2000

8/12/1981
IBM PC Intro

2005

6/29/2007
iPhone released (USA)

2010

6/1/2000
AMD Athlon Thunderbird
(1 GHz X86)

2012

4/12/2012
Google Glass announced

2014

4/12/2012
Google Glass announced

2014

9/19/2014
***TODAY***
CLIENTS

• Diversity
  • Phones, tablets, laptops
  • Wearables
  • IOT

• Cloud integration
  • The internet at your fingertips
DRIVERS FOR CLIENT COMPUTER ARCHITECTURE?

- PPA (performance, power, area)
- Specialization (fixed function)
  - Phones Phone
  - Phones Interact
  - Cameras camera
  - Refrigerators refrigerate
- Rapid time-to-market
CLOUD

- PPA again the big driver
  - Aka, perf/watt/$
- Virtualization
- Security
- Storage Architecture
  - Flash, cheap disks, IOPs
  - SANs
- Connectivity
  - 5G, WiFi, BlueTooth, NFC
  - Copper, Fiber
A NEW WORLD COMING?

The Rise of the Mobile GPU & Connectivity
DISCRETE GPU MARKET

Graphics shipment 1981 to present (M units)

- Desktop GPUs
- Notebook GPUs
- Desktop PCs
- Notebook PCs
- TOTAL Graphics (M units) 32 Years CAGR 19.72%
- TOTAL PCs (M units) 32 yr. CAGR 18.4%

Tailing Off
In 2013, an estimated 1.2B+ mobile GPUs shipped
- ~200M tablets
- ~1B smart phones
Continues to grow, but saturation on the horizon?

Trend:
- Discrete GPU decreasing
- Mobile is growing

Source: IDC Worldwide Mobile Phone Tracker, January 27, 2014
WW INTERNET TRAFFIC

- Internet traffic growth rate is staggering
  - 2013 total traffic estimated at \(~51.1\) EB per month
  - By 2018, more than double
  - 2014 per person smartphone traffic at \(~8\) GB per month on WiFi and \(~3\) GB per month on broadband
  - 2014 per person tablet traffic less, but still \(~3.2\) GB per month WiFi and \(~0.4\) GB per month broadband
WHERE ARE WE HEADED FROM A HW PERSPECTIVE?…

- Enormous quantity of GPUs
- Large amount of interconnectivity
- Better I/O
The OpenGL ES 2.0 pipeline

GPU Pipelines

A BRIEF REVIEW OF GPU TECH
MOBILE GPU PIPELINE ARCHITECTURES

*Tile-based immediate mode rendering (TBIMR)*

*Tile-based deferred rendering (TBDR)*

IA = input assembler  
VS = vertex shader  
CCV = cull, clip, viewport transform  
RS = rasterization, setup  
PS = pixel shader  
ROP = raster operations (blend)
**TBDR W/ HSR**

- **HSR** = *hidden surface removal*
  - Sort all objects across each projection ray
    - Use tiling to reduce data set size
  - Only nearest opaque and closer transparent objects need to be drawn
  - Remaining fragments can be killed => not drawn
## MOBILE GPU LANDSCAPE

<table>
<thead>
<tr>
<th>Company</th>
<th>Product</th>
<th>Pipeline</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM</td>
<td>Mali</td>
<td>TBIMR</td>
<td>Unified shader, 2-4 math pipes per core</td>
</tr>
<tr>
<td>Imagination</td>
<td>PowerVR</td>
<td>TBDR/HSR</td>
<td>Latest is Series6XT. Unified shader. DX11 support</td>
</tr>
<tr>
<td>Qualcomm</td>
<td>Adreno</td>
<td>FlexRender</td>
<td>Unified shader. “FlexRender” = automatic switching between direct render (IMR) and tile-based deferred rendering (TBDR).</td>
</tr>
<tr>
<td>NVIDIA</td>
<td>Tegra</td>
<td>TBDR &amp;</td>
<td>Evolution:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TBIMR</td>
<td>• Tegra 1/2/3/4: non-unified TBDR architecture</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Logan: Kepler-based GPU, TBIMR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Parker: Maxwell-based GPU, TBIMR</td>
</tr>
<tr>
<td>Vivante</td>
<td>ScalarMorphic</td>
<td>IMR</td>
<td>Unified Shader.</td>
</tr>
<tr>
<td>Intel</td>
<td>Gen</td>
<td>Atom</td>
<td>IMR</td>
</tr>
<tr>
<td>AMD</td>
<td>Radeon</td>
<td>IMR</td>
<td>Mobile R9 M2xx series (2014)</td>
</tr>
</tbody>
</table>
A PATH TO A BETTER MOBILE GPU?
[PART 1]
WHAT IS IMPORTANT?

- More with less
- Better user experience
PARALLELISM

• Parallel vs. Sequential
  • Parallel $\rightarrow$ independence
  • Sequential $\rightarrow$ dependence

• Three fundamental forms of parallelism
  • Spatial: executing operations between threads at the same time
  • Temporal: executing operations between threads at the same place
  • ILP: executing operations from within the same thread in parallel

• Fundamental differences between ILP-only machines and massive TLP-ILP machines
  • CPUs vs. GPUs
THROUGHPUT VS. LATENCY

- Throughput = rate at which operations complete
- Latency = time it takes to complete an operation or set of operations
- CPUs versus GPUs
  - In CPUs, the primary objective is low latency
  - In GPUs, the primary objective is high throughput
- CPUs versus GPUs
  - In an application suitable for CPUs, we assume a low degree of TLP
  - In an application suitable for GPUs, we assume a high degree of TLP
GPU PERFORMANCE

- Supply and demand:

\[ \hat{S} \geq \lambda \hat{D} \]

(“limiter equation”)

- Lambda (\(\lambda\)) is throughput
- Supply examples:
  - FP BW (flops/clock)
  - Texture BW (quads/clock)
  - Memory BW (bytes/clock)
- Demand density examples:
  - FP ops per shader
  - Sample ops per shader
POWER EFFICIENCY

• Performance = power efficiency

• Two types of efficiency:
  • “perf@watts”:
    The ability to deliver maximum performance
  • “watts@perf”:
    The ability to deliver maximum battery life at a minimum required performance
WHAT IS EFFICIENCY?

- **Perf @ Watts**
  - *Maximum performance at some power limit*
  - Limits:
    - electrical (Pidd)
    - die temp (Tj)
    - skin temp (Tcase)
    - battery life (Pbat)

- **Watts @ Perf**
  - *Minimum power at constant performance*
  - Example: deliver 60 frames/sec at lowest power
ENERGY REDUCTION TECHNIQUES

- Work Reduction
- Memory Avoidance
- Memory BW Reduction
- Memory Access Management
WORK REDUCTION

- Pixel shaders in ES games ~95% of the shader load
  - A pixel shader killed is raw power savings
  - HSR can kill 30-50% of the shader threads

- Geometry in DX11 a problem
  - Unigine Heaven ~10M Tri/frame
  - Can be up to 70% of shader workload

- Inter-frame work reduction?
RELATIVE ACCESS ENERGY COSTS

- LPDDR
- WIO1
- SRAM
- Small RF
- SP FMA

- Energy/byte
- Energy/ope
MEMORY AVOIDANCE

• Memory power a problem
  • LPDDR ~100 pJ/byte
    (100 mW @ 1 GB/sec)
  • WIO1 ~24 pJ/byte
    (24 mW @ 1 GB/sec)
  • On-chip SRAM ~0.6 pJ/byte
    (0.6 mW @ 1 GB/sec)

• Reduction in working set for non-essential traffic (i.e., not texture, attribute, command, or render target)
  • Rematerialize? (computation vs. BW)
  • Scheduling to reduce lifetimes?
MEMORY BW REDUCTION

- Texture compression (RD)
  - Better compression?
  - Tessellation use of textures?

- Tile compression (WT)
  - TB-based signature checking
  - Lossless compression

- Attribute compression (RD)
  - Reduce stream BW
MEMORY ACCESS MANAGEMENT

- SOC memory architecture
  - Blood rivals (antagonists)
  - Effect of CPU/GPU traffic on Memory Controller (MC)
    - Intelligent page open/close management
    - Balance latency vs. BW

- Mismanaging DRAM results in both performance loss AND extra energy – double whammy
A better user experience…

A PATH TO A BETTER MOBILE GPU? [PART 2]
ISO 9241-210[1] defines user experience as "a person’s perceptions and responses that result from the use or anticipated use of a product, system or service". - Wikipedia
APPLICATION: NAVIGATION

Hercules
Rises: 6:03 a.m.  Sets: 9:25 p.m.
A large constellation representing
the mythological hero

U.S. Capitol complex
0.7 miles
... Construction of the Capitol began in 1793.
When built, it was ...

Sanphan restaurant
★ ★ ★ ★ ★ 41 reviews
Thai  $$

Car locator

Eastern Market
580 feet
Turn right on 7th St.

Gas station
550 feet
Unleaded
$3.19

Robbery
270 feet west
18 days ago

CORONA
BOREALIS

HERCULES
APPLICATION: FACE RECOGNITION
APPLICATION: TELEPRESENCE

http://www.youtube.com/watch?feature=player_detailpage&v=Nzi0sm81tP4

APPLICATION: VIRTUAL COMPUTER
THE UX OPPORTUNITY

• Killer apps will be integration of:
  • AR/MR technology
  • Big Data operations

• Subject to:
  • Real-time constraints
  • Parallelization on a massive scale
Making a better UX

FUTURE MOBILE TECH CHALLENGES?
KEY CHALLENGES

• I/O:
  • AR Headsets
  • Environment Imaging
  • IOT integration

• Computational:
  • API Improvements
  • Cloud-device integration
AR HEADSETS

• Google Glass is pretty cool, but…

• Better imaging
  • Stereo/Light field
  • HD → UHD
  • Speed

• More sensors

• Wireless power?

• Fashion/ubiquity
ENVIRONMENT IMAGING

- For telepresence, headset camera is insufficient
- Need “environment cameras”
- Lots of privacy concerns
- Localizing environment to a client?
API IMPROVEMENTS

• Today’s APIs are power inefficient

• Needed:
  • Hints
  • State-less rendering
    • API commands supply state with action
  • Frame-less rendering
    • Compositing deferred and on-demand
  • Hierarchical geometry
    • Deferred detail
CLOUD-DEVICE INTEGRATION

• SW Challenge:
  • Making cloud queries easier
  • Utilizing the parallelism of the cloud

• Ultimate challenge:
  • The “network GPU”
  • Analogously extend the GPU model to network scale
  • $10^9$ GPUs $\rightarrow 10^{21}$ FLOPs?
SUMMARY

• Computing has changed our world and will continue to do so
• Mobile computing, in particular graphics, is growing rapidly and becoming ubiquitous
• Tomorrow’s machines:
  • Ever improving efficiency
  • Integrated visual UX
  • Tied to the cloud
• Challenges remain to make this a reality
• Exciting prospects…