Cloud Gaming & Application Delivery with NVIDIA GRID Technologies

Franck DIARD, Ph.D.
GRID Architect, NVIDIA
What is GRID?

- Using efficient GPUS in efficient servers

What is Streaming?

- Transporting pixels to a client through the network
- Interactive graphics = Low latency!!!
Content for Cloud Graphics

- Native games
  - Windows DirectX / OpenGL
  - Linux OpenGL
- VM Players
  - Android Games
    - VirtualBox + Android x86
    - Flash player
- New Engines
  - Shiva 3D
- Workstation/VDI
  - VMWare, CITRIX, NICE
- Virtual set top box/transcoding
  - Decode
  - Rendering STB UI
  - Recode
Cloud Gaming Advantages

For Publishers

- No Piracy
- Precise Accounting
- High performance HW
- Controlled SW
- Game Updates

For Gamers

- Mobility
- Ubiquity
- Instant Play
GRID Technologies

- GRID SDK:
  - Set of HW and SW to stream GPU accelerated graphics from the cloud or on premises
- HW accelerated rendering
- Video compression acceleration
- GPU virtualization
  - Different architectures for different usage models
- A team of NVIDIA SW and HW experts
  - Interacting with a very rich middleware and HW solutions ecosystem
Strategies

- **GPU desktop/application streaming middleware**
  - Use a middleware stack that uses GRID SDK
  - Build your own using the GRID SDK

- **Cloud servers**
  - Use a GRID powered IAAS service
    - AWS
  - Build your own
    - GRID HW components
    - Server design templates
Building a Streaming Middleware
GRID SDK Components

Server Side on GRID HW
- NvFBC
  - Low Latency
  - Desktop
  - Capture
- NvIFR
  - Low Latency
  - Render Target
  - Capture
- NvENC
  - Low latency
  - Hardware Encoder

Client Side on NV
- Low Latency Decode
  - on GeForce & Tegra

Sample Code (NvIFR & NvFBC)
- GRID Programming Guide
- GRID Server Setup Guide
- Amazon G2 Getting Started Guide
GTC On-Demand

GPU computing is a transformational force in high performance computing and is enabling developers, engineers, programmers and researchers across a myriad of industry verticals, as well as academia to accelerate research and mission critical applications. See our featured sessions highlighting some of our best talks or dive head-long into the many other keynotes, technical sessions, presentations, research posters, webinars and tutorials we make available to you at any time on GTC On-Demand.

REMOTE GRAPHICS & CLOUD-BASED GRAPHICS

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerating Cloud Graphics</td>
<td>Streaming:</td>
</tr>
<tr>
<td>Franck Diard (NVIDIA)</td>
<td>&gt; Watch Now</td>
</tr>
</tbody>
</table>

Franck Diard, Chief Software Architect at NVIDIA, will talk about the technologies behind GRID and how you can integrate them into your cloud products. The audience will learn about the key components, which allow optimal capture, efficient comp ... Read More

Keywords: Remote Graphics & Cloud-Based Graphics, Media & Entertainment, GTC 2013 - ID S3543

CLOUD VISUALIZATION

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerating Cloud Graphics</td>
<td>Streaming:</td>
</tr>
<tr>
<td>Franck Diard (NVIDIA)</td>
<td>&gt; Watch Now</td>
</tr>
</tbody>
</table>

A new NVIDIA SDK provides access to a class of key components which allow optimal capture, compression, streaming and low latency display of high performance games from the cloud. We demonstrate how all these components fit together to deliver a ... Read More

Keywords: Cloud Visualization, GTC 2012 - ID S0627
2 Different APIs: NvFBC and NvIFR

- For VERY different things
  - NvFBC is an asynchronous display grabber, wholesale capture, orthogonal to any apps or graphics API
  - NvIFR is a transfer of the rendered frame of the currently bound render target, per 3D context

- NvFBC capture only what is seen on a display, nothing else
  - one grabber per virtual display

- NvIFR is only used in a DX9, DX1x and OpenGL context
  - Does not matter if rendered graphics are on the screen or not
  - As many sessions as rendering contexts
Common Features to NvFBC and NvIFR

- HW H264 encoding
  - Fast path between FB and NVEnc
- DMA readback to page locked system memory
  - For CPU codecs
  - Various format conversions (YUV, YUV444, RGB, planar)
  - Completely DMA
- CUDA mapping
  - Other codecs
NvFBC Usage Case

Brute force "whole screen grabbing" but has some advantages
Easier onboarding of apps, no process injection
Uses a simple fullscreen grabber in different process
NvIFR Usage Case

- Game/App is injected/hooked by streaming middleware or uses the SDK natively (AWS AppStream, Shiva)

- Used along with lightweight virtualization shims layers/application delivery
NVEnc

Fast, asynchronous to 3D engine
Only uses 2 watts per GPU
High quality
New Feature of the GRID SDK/HW

YUV444 vs. YUV420

420 coding

444 coding
Running Several Applications per GPU
GPU Virtualization

- Application shimming
  - 1 OS = n GPUs = m apps
  - Linux, Windows “baremetal”
  - No static partition of resources
  - No CPU tax

- NMOS
  - n OS = n GPUs = n desktops

- vGPU
  - n OS = m GPUs = n desktops (n>>m)
Architectures: No ‘Virtualization’

- But Poor man’s GPU virtualization can be used
Multiple DX/OpenGL Contexts
Using API Injection and NvIFR
Increasing CCU with Shimming/Sandboxing

- Allows launching multiple instances
- File system isolation
  - Persistence
- Requires forcing fullscreen games off the screen

1 single OS, n sandboxes
High CCU for Casual DX Games

- SF4: Xeon E5 with 3 K340s: 60+ sessions
GRID SDK/HW CCU Benefit

- Portal - SFR
- Rabbits Go Home
- Pure
- Lego Batman 2
- Lego Pirates
- Mirror's Edge
- Prince of Persia
- Assassin's Creed 2
- Farming Simulator 2013

Dual Xeon E5-2670

Game Streams

# of Concurrent 1280x720@30fps Game Streams

- 2 x K520-H.264
- 2 x K340-H.264
- 2 x Quadro 4000-MPEG-4
Performance, Load Balancing

- **Resource Monitoring**
  - CPU load and system memory usage using OS API
  - GPU load and FB usage through NVML

- **CPU Affinity**
  - Undo application settings
  - Mostly locking core for desktop perf

- **NUMA**
  - Minimize QPI/HT traffic
  - Resource placement
Architectures: XEN NMOS

- The platform is virtualized, NOT the GPUs
- To be used with NvFBC
Architectures: XEN vGPU, all virtualized
# vGPU Session

## TALK

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Details</th>
</tr>
</thead>
</table>
| **S4725 - Delivering High-Performance Remote Graphics with NVIDIA GRID Virtual GPU** | **Day:** Tuesday, 03/25  
**Time:** 14:00 - 14:50  
**Location:** Room 210F |

[Andy Currid](#) (System Architect, Distinguished Engineer, NVIDIA)

Learn how to deploy and optimize high-performance remote graphics applications using NVIDIA GRID Virtual GPU. This session will include an architectural overview of GRID Virtual GPU, which provides true hardware virtualization and sharing of the GPU between multiple virtual machines, a walkthrough of Virtual GPU setup on Citrix XenServer with remote graphics, and examples of how to tune the configuration for optimum remote graphics performance.
GRID SDK, Client Side
Building Optimized Clients with GRID SDK

- Not to be overlooked
- GRID Decode API is designed for Low latency (1 frame)
- Power efficient using the H.264 hardware decoder

GRID SDK Examples:
  - Windows PC
    - Based on **DX9CUDA\GPUDecode**
  - NVIDIA Shield (Android)
    - Based on **TegraH264HWDecode**
GRID HW Platforms
## GRID Gaming Boards

<table>
<thead>
<tr>
<th>Product Name</th>
<th>GRID K340</th>
<th>GRID K520</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPUs</strong></td>
<td>4 x GK107</td>
<td>2 x GK104</td>
</tr>
<tr>
<td><strong>Total shader (SMs)</strong></td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td><strong>Core Clocks</strong></td>
<td>900 MHz</td>
<td>800 MHz</td>
</tr>
<tr>
<td><strong>Memory Size</strong></td>
<td>4GB</td>
<td>8GB</td>
</tr>
<tr>
<td><strong>720p30 Encodes</strong></td>
<td>~40</td>
<td>~20</td>
</tr>
<tr>
<td><strong>1080p30 Encodes</strong></td>
<td>~16</td>
<td>~8</td>
</tr>
<tr>
<td><strong>Driver Support</strong></td>
<td>GeForce Drivers and Game Profiles</td>
<td>GeForce Drivers and Game Profiles</td>
</tr>
<tr>
<td><strong>Target Apps</strong></td>
<td>High Density Gaming</td>
<td>High Performance Gaming</td>
</tr>
</tbody>
</table>

- Professional GRID
- K1 and K2
- Enabled with vGPU
- Certified with Pro apps

1 Number of users depends on application and screen resolution
Building GRID Gaming Servers

- **Dual Xeon E5**
  - 2x10 core 2.5 GHz CPUs
  - SuperMicro, ASUS
  - Up to 5 boards = 20 GPUs

- **Xeon E3 systems**
  - 4x 4-core 3.5 GHz CPUs
  - 2 boards = 8 GPUs
  - Cirrascale, CARRI/GIGABYTE

- **Xen, baremetal**
Amazon Web Services with G2

- A GPU accelerated VM on the AWS
  - Amazon G2 instance: **g2.2xlarge**
    - 1 NVIDIA GRID GK104 (GeForce GTX 670)
    - 4-Core CPUs/15GB Memory/ 60GB Storage
  - Rent Amazon G2 instances for $0.78-$0.82 per hour. Spin up/down instances based on traffic demand.
  - BIG bandwidth to internet and storage
Cloud Streaming Middleware
A Rich GRID Gaming Middleware Ecosystem

- SFR France/G-Cluster
- Bouygues France/Playcast
- Orange France/G-Cluster
- LG Uplus Korea/Ubitus
GPU Desktop Streaming Middleware on AWS

- Windows Server on AWS G2, using GRID SDK natively
- TeamViewer
  - Feature rich remoting software
  - Load it into your instance by simple install
- Scalable Graphics
  - Published AMI, Dedicated Client
  - http://www.scalablegraphics.com/aws/
MainFrame2: Application Delivery Middleware

- **MainFrame2**
  - Automatic ISV app onboarding
  - HW/SW provisioning on AWS
  - Streaming

- **GRID HW accelerated for rendering and compression**

MainFrame2 is in private beta. Sign up now for early access and exclusive updates.
TransGaming.com

- Running DX Games on top of Linux
- Solving MS licensing costs
- High CCU
Android Gaming Streaming

- **Android-x86 in VirtualBox**
  - Thin provisioning
  - Mapping VirtualBox’s FB to NvIFR/NVENC

- **Gaming**
  - API intercept of OpenGLES
  - Rendered and encoded in Host

1 single OS, n instances of headless VirtualBox
vSTB: Fixed Streaming Application

- Big volume market
- Up to 200 streams per server
- Linux based
- Single application to shim: Chrome/WebKit
  - Decode
  - Decorate with GUI
  - Stream

1 single OS, n instances of offscreen Chrome
Shared World Engines on GRID Powered Servers

- Shiva3D engine
Resources


- GRID-devtech-support@nvidia.com
- GRID General Manager
  - phil.eisler@nvidia.com

- Thanks!