GPU Virtualization on VMware's Hosted I/O Architecture

Micah Dowty
Jeremy Sugerman

USENIX WIOV 2008

Contents

- GPUs are hard
- But GPU virtualization is worth the trouble
- How to virtualize a GPU?
- VMware's virtual GPU
- Conclusions

- In the paper:
  - Details on our implementation
  - Benchmarks, analysis

What is a GPU, anyway?

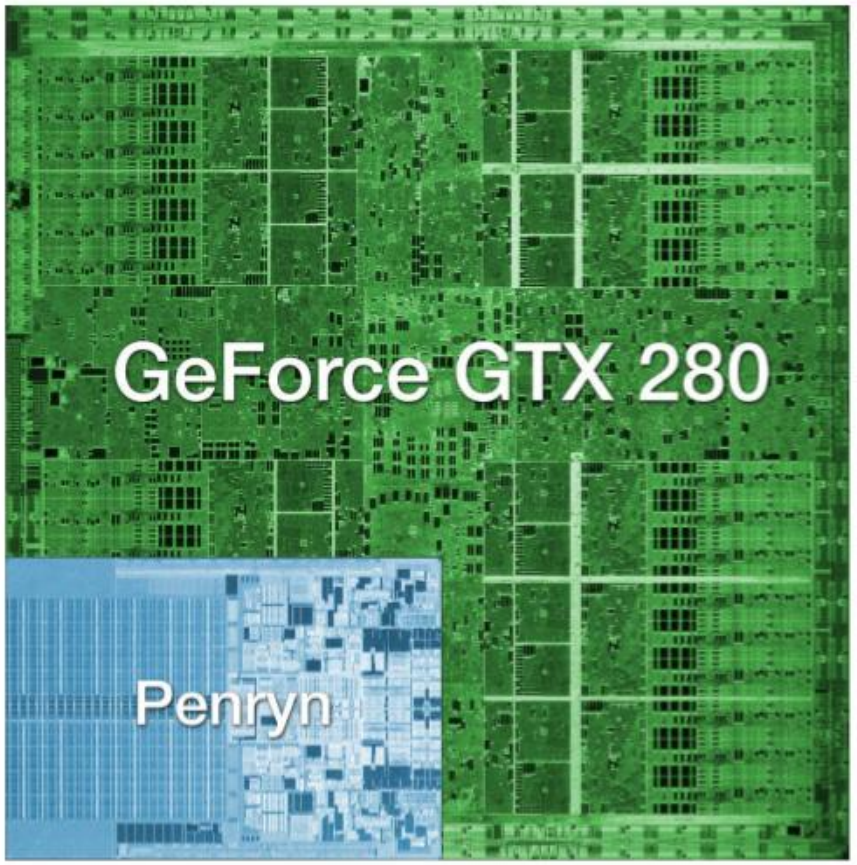
- Video playback, 2D graphics, drawing triangles and rectangles and lines...

- Computation.

How much computation?

NVIDIA GeForce GTX 280:

1.4 billion transistors

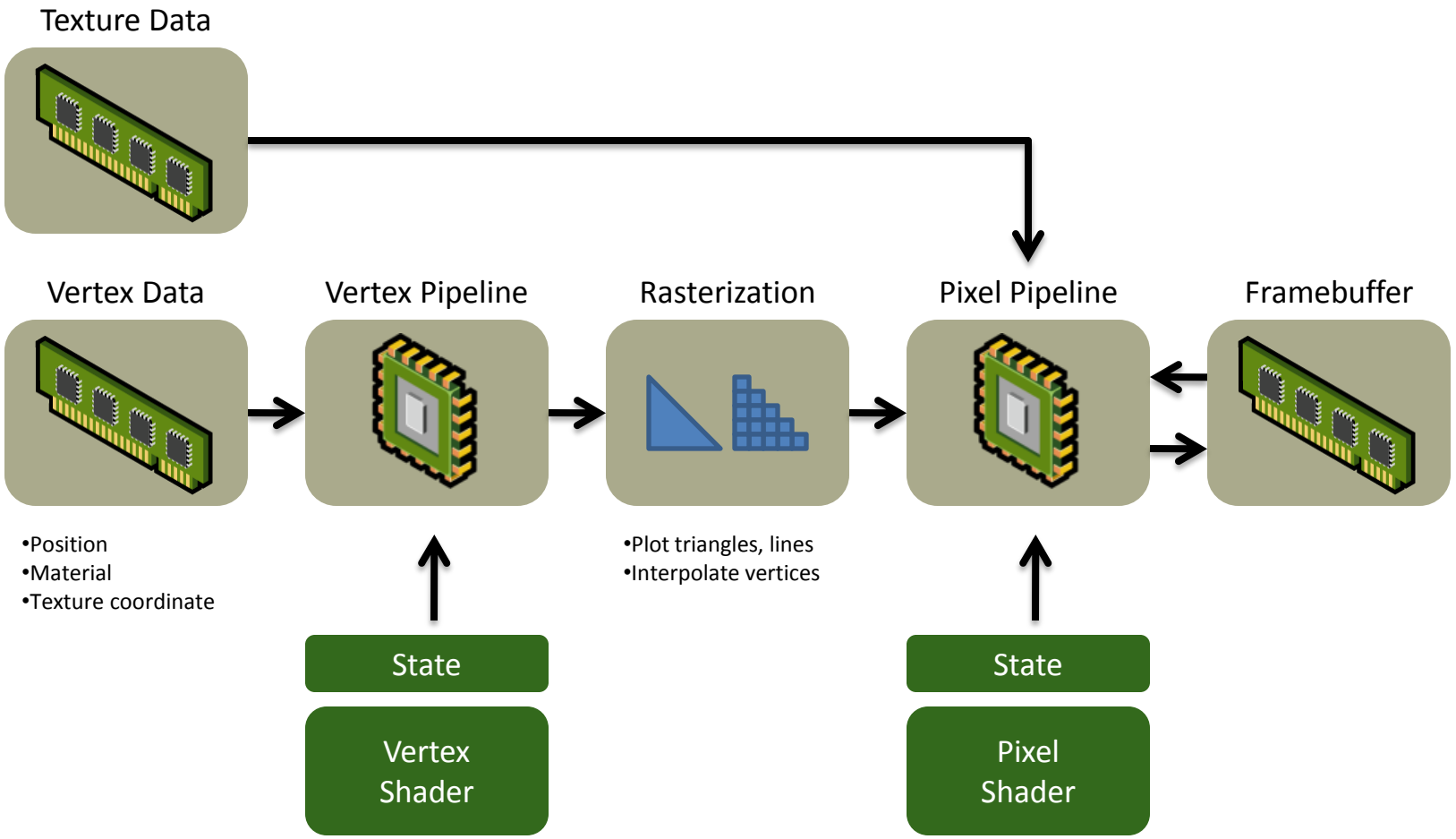


Intel Core 2 Duo:

291 million transistors

Source: AnandTech review of NVidia GT200

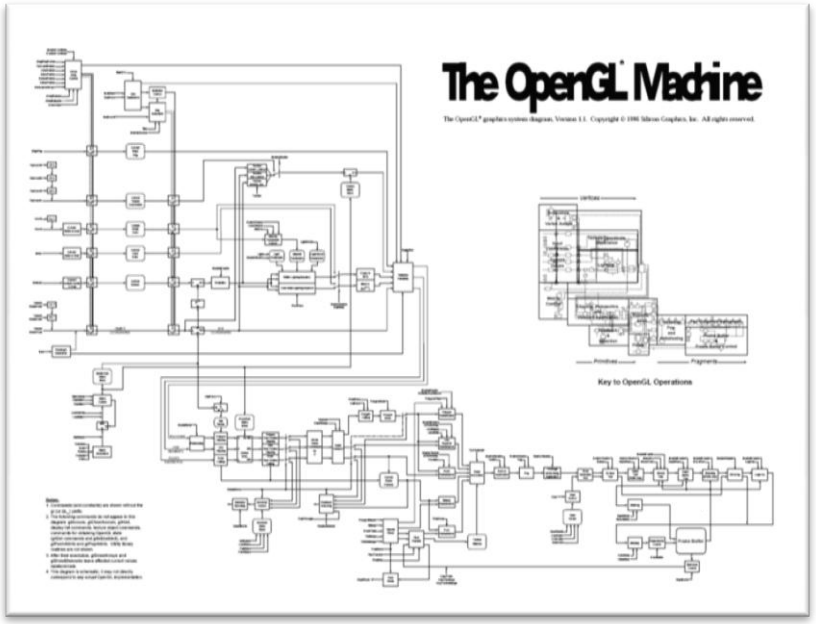
Programmable 3D Pipeline



(State of the art circa 2002...)

Unique challenges

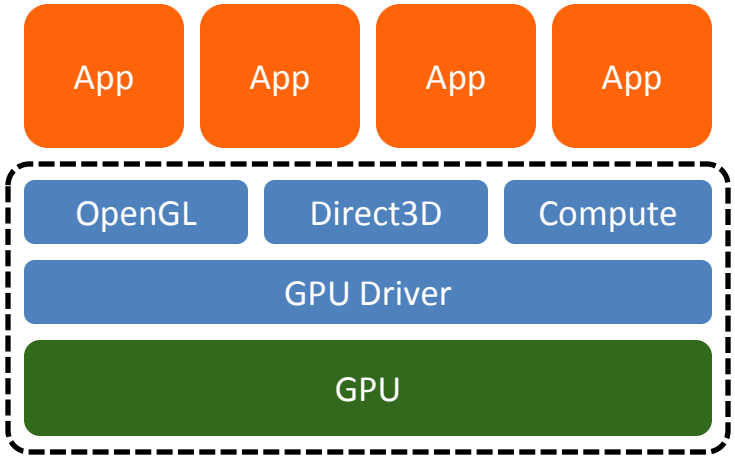
- API
  - Not quite read(), write(), select()...
  - Multiple competing APIs
  - Hundreds of entry points



- Programmable
  - Every GPU driver is also a compiler
  - Each API includes a language spec

Unique challenges

- Hardware specs
  - Diverse, changes frequently
  - Closely guarded secret*
  - Speed vs. portability

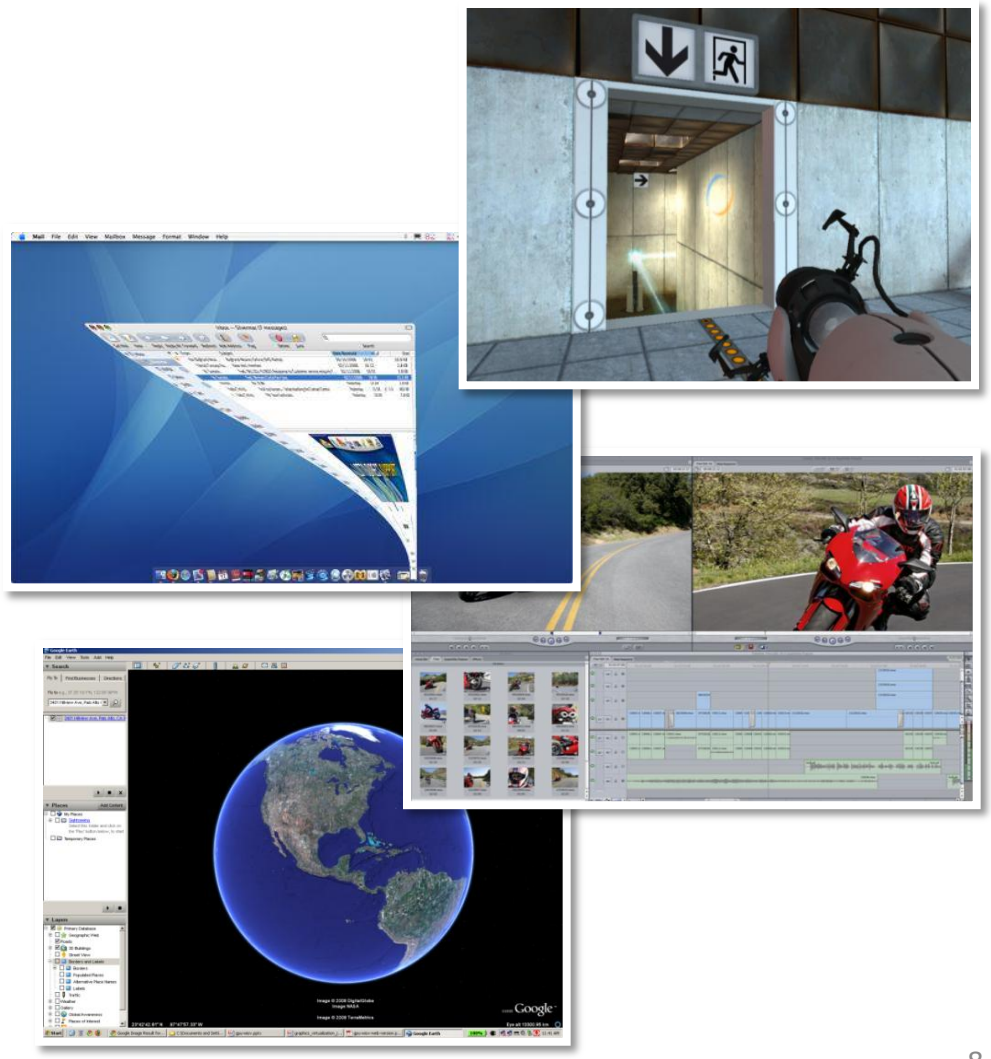


- Hardware state
  - Up to gigabytes of data
  - Highly device-specific format
  - In-progress DMA and computation

* With a few notable exceptions.

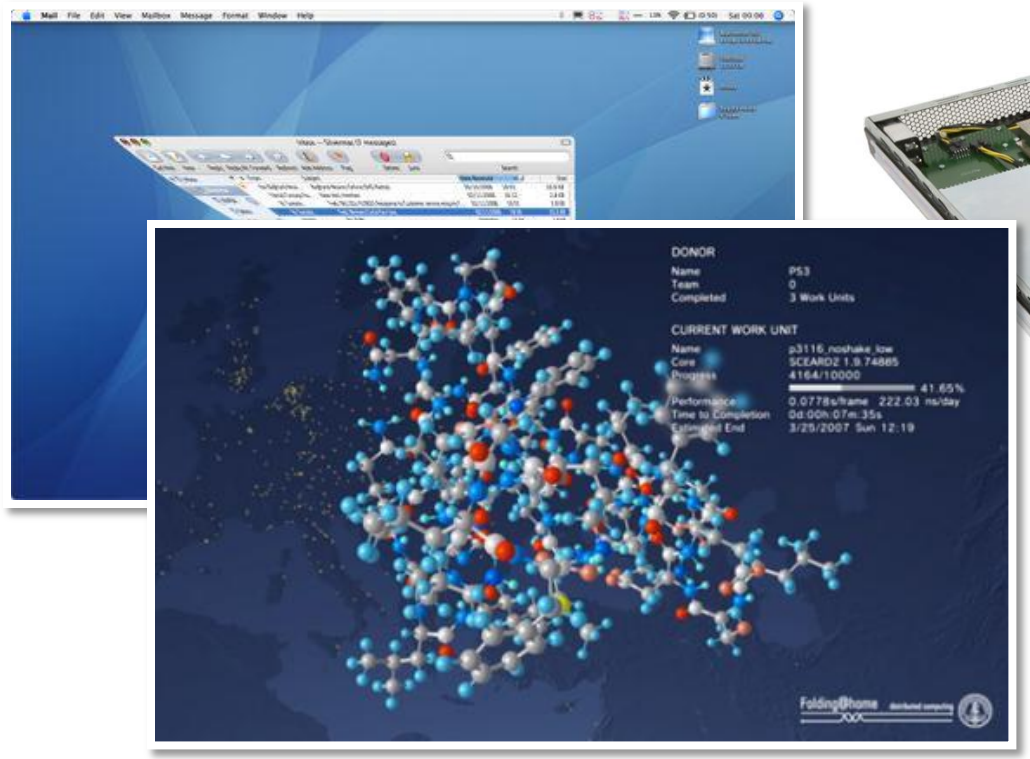
What are GPUs good for?

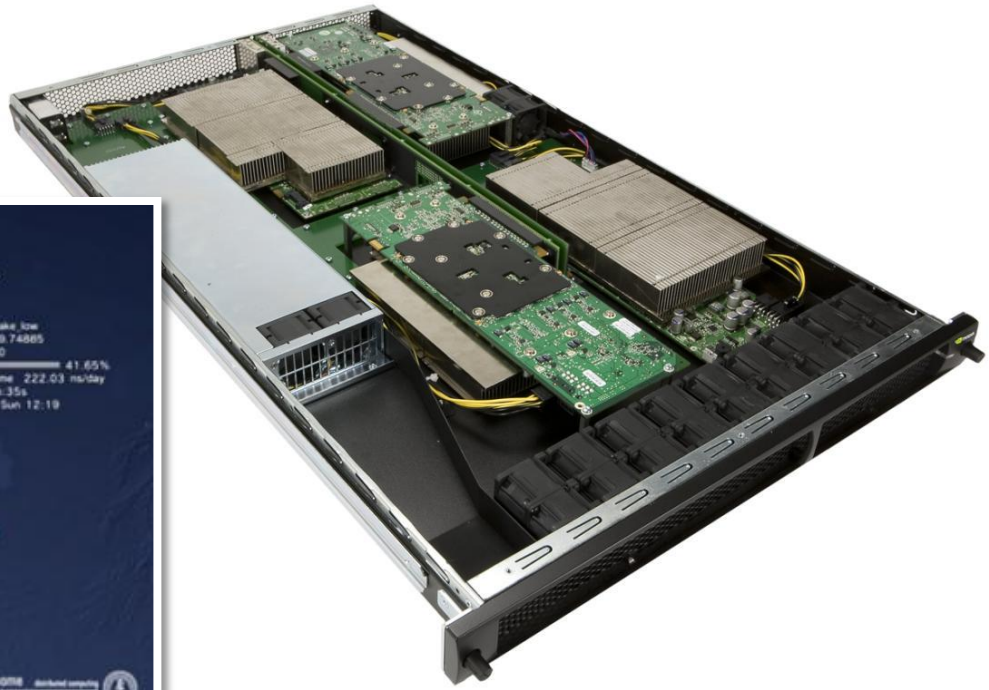
- Desktop Apps
  - Entertainment
  - CAD
  - Multimedia
  - Productivity
- Desktop GUIs
  - Quartz Extreme
  - Vista Aero
  - Compiz



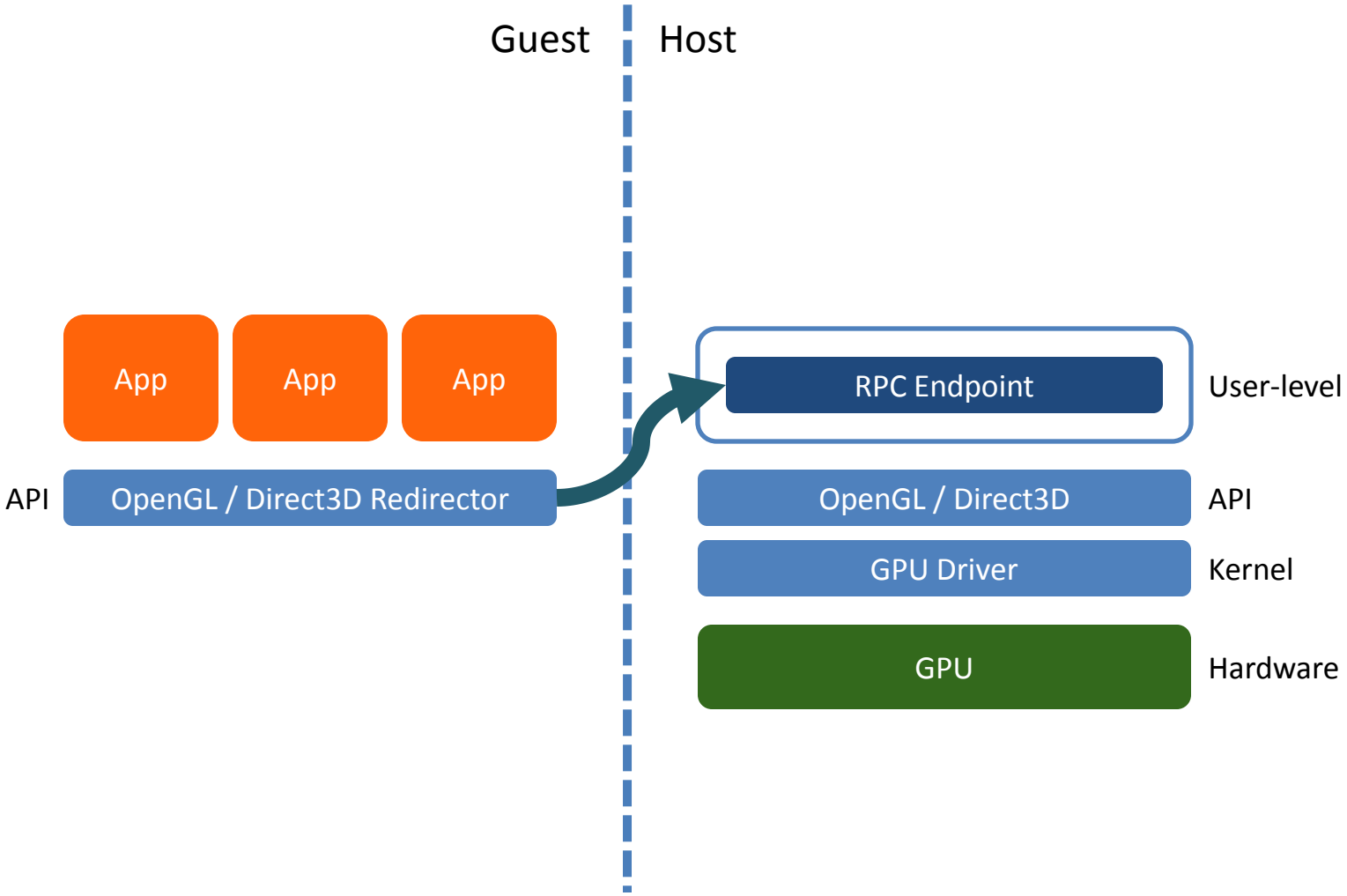
GPUs in the Data Center

- Server-hosted Desktops
- GPGPU

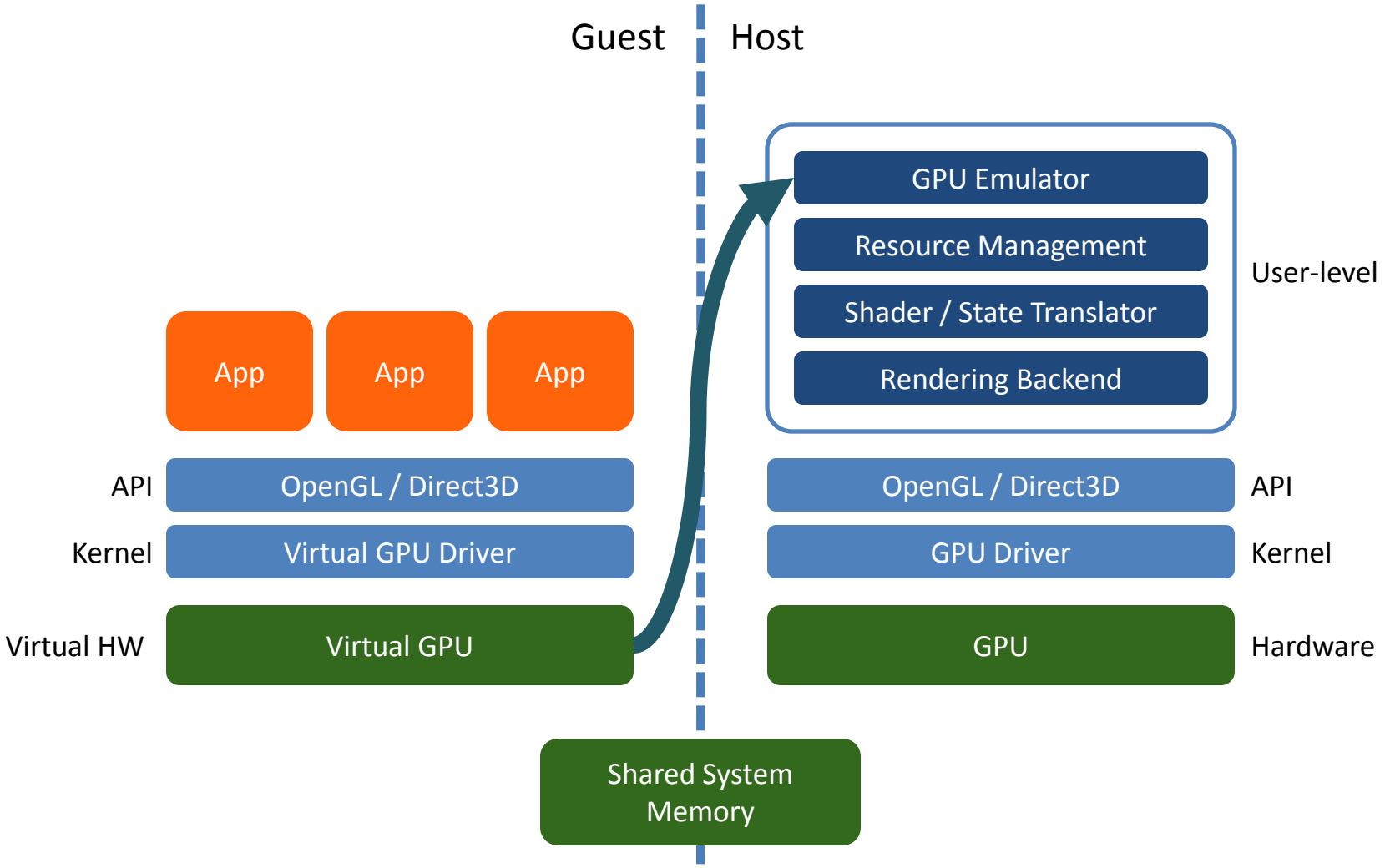




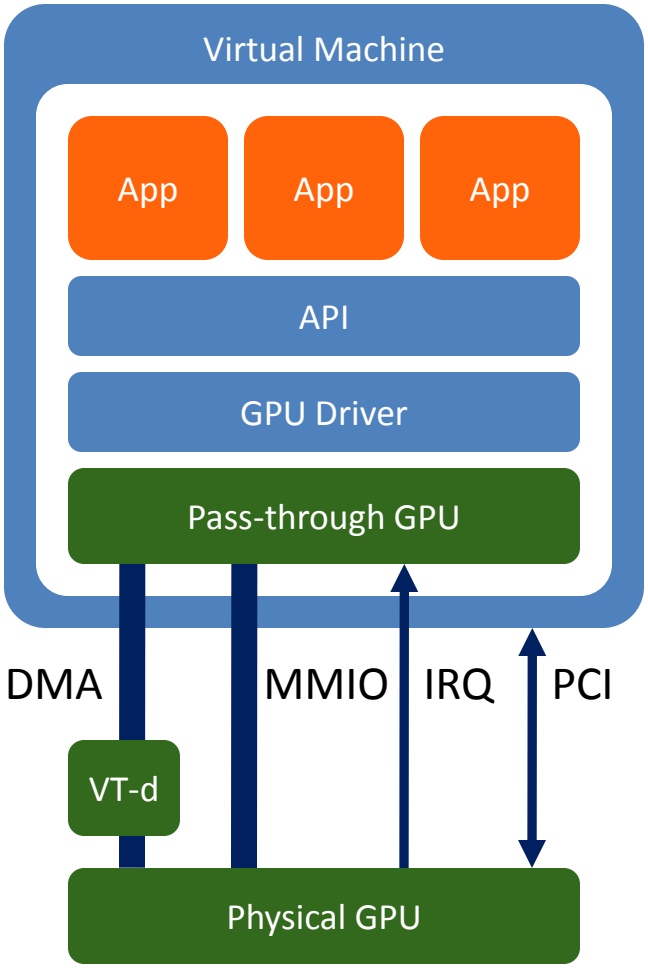
API Remoting



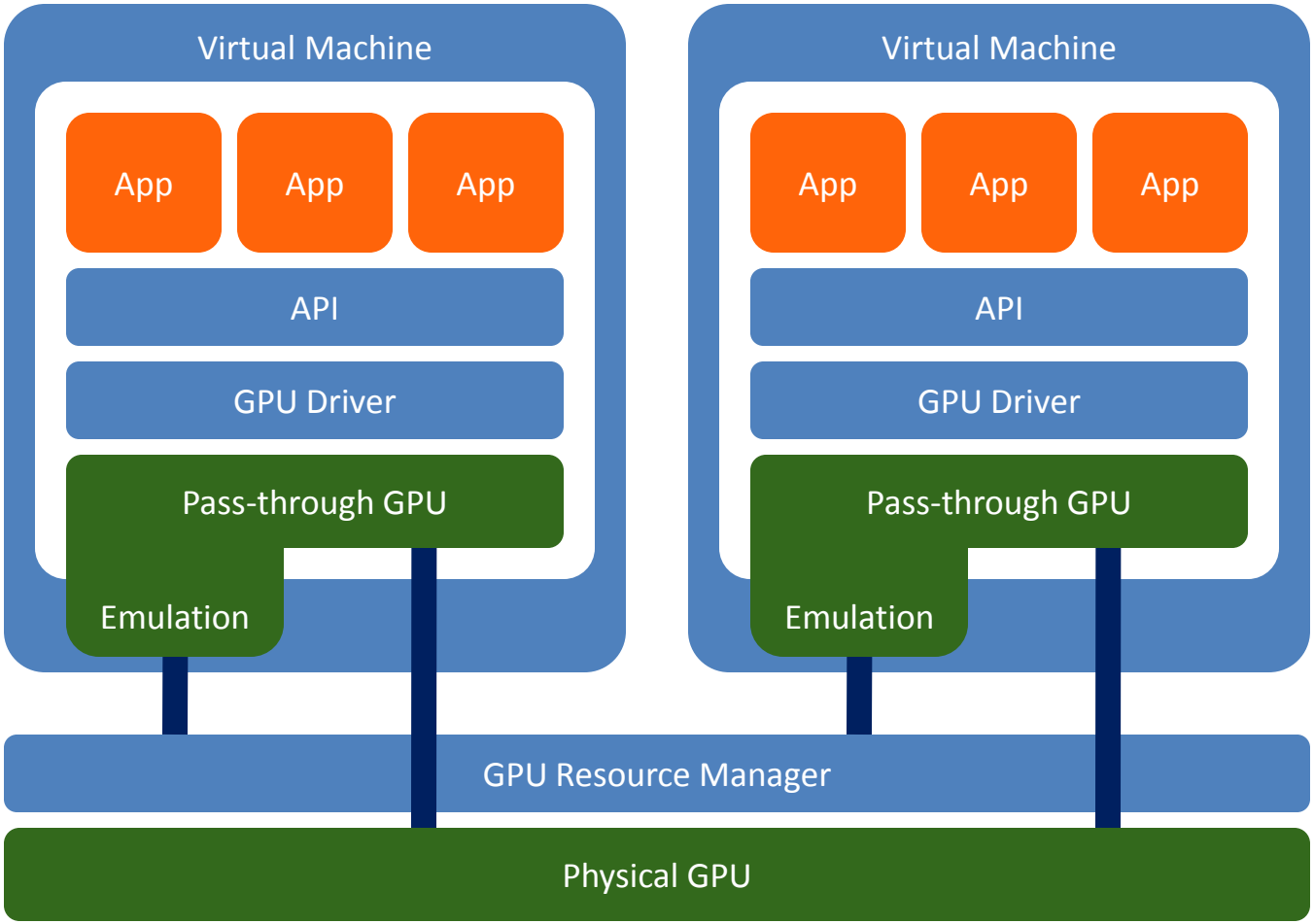
Device Emulation



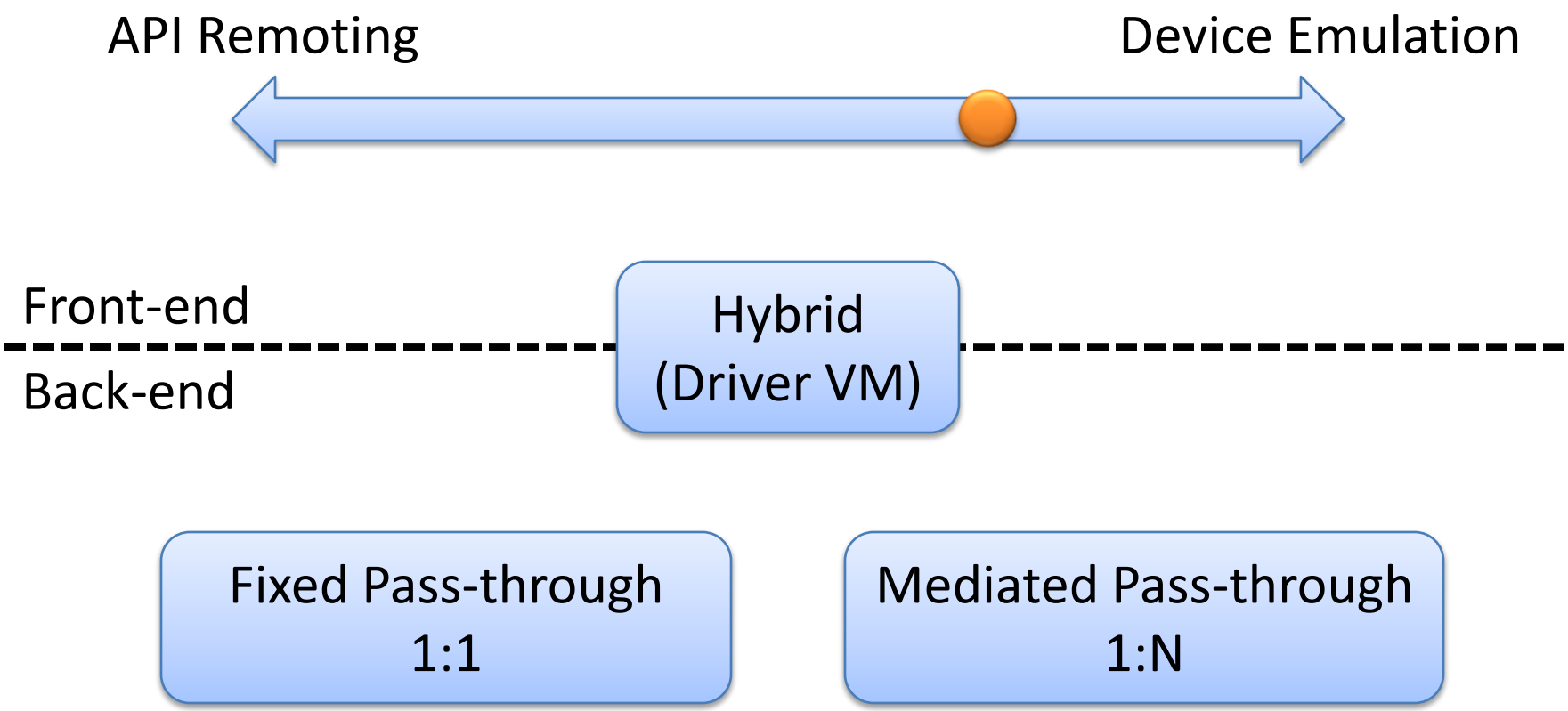
Fixed pass-through



Mediated pass-through

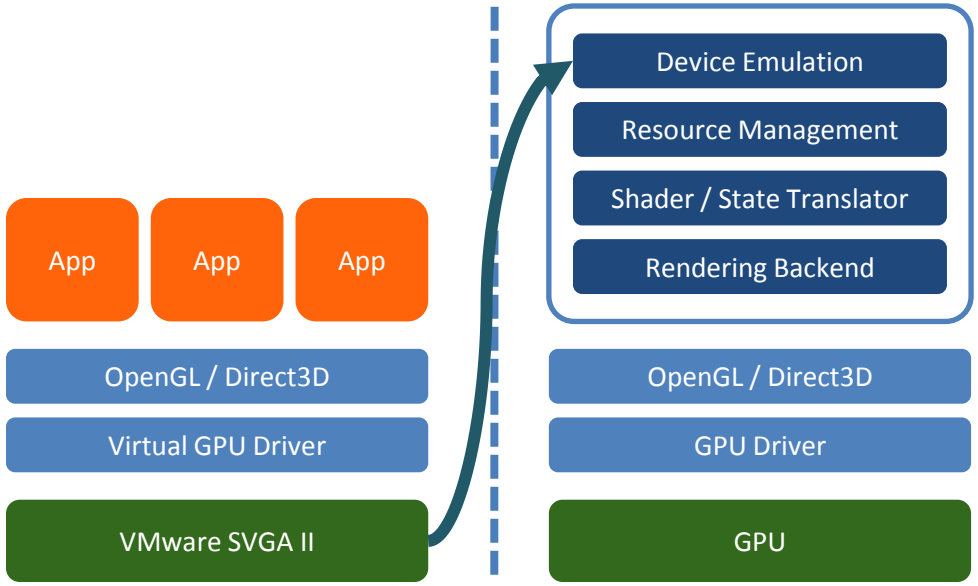


GPU Virtualization Taxonomy



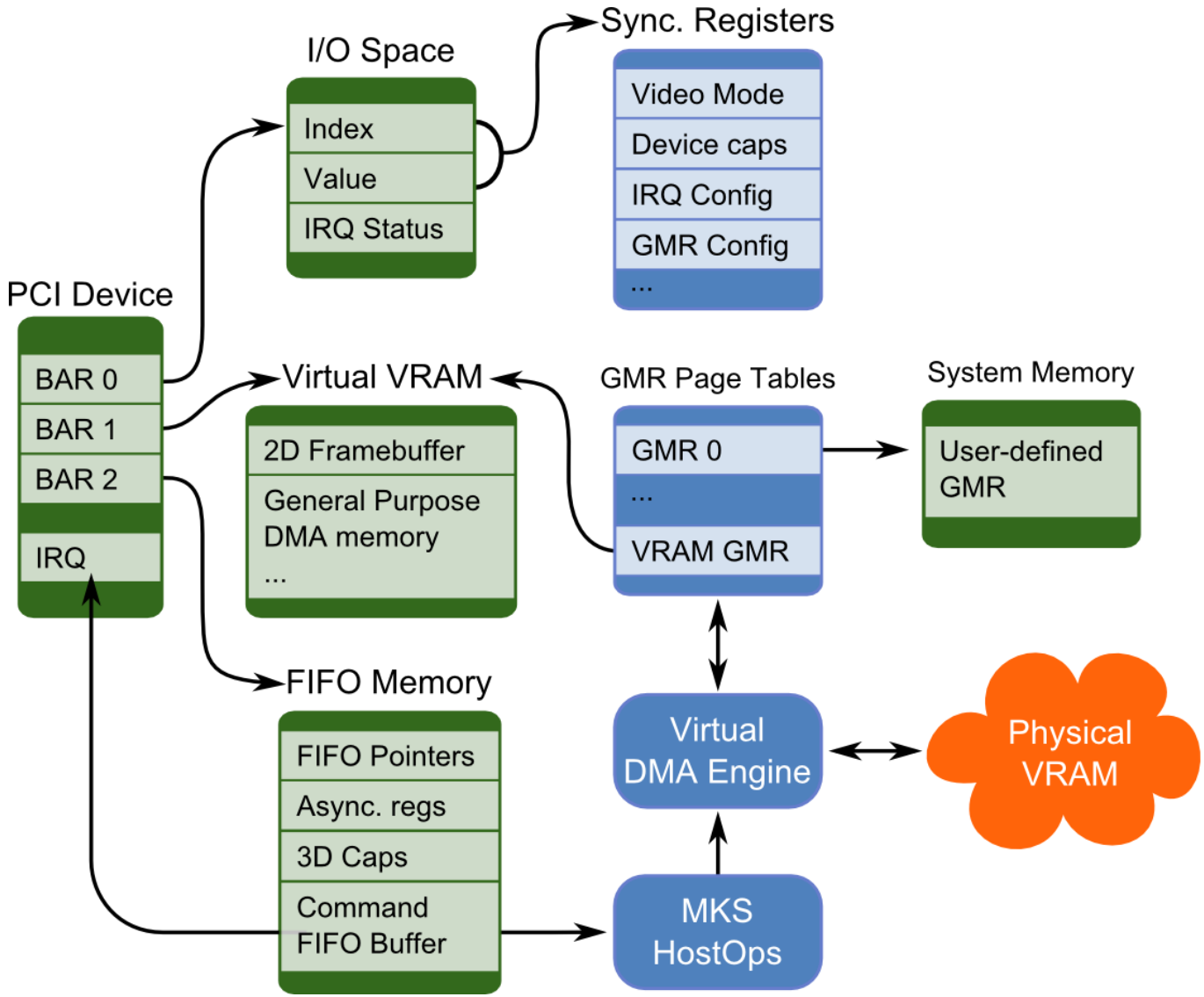
VMware's Virtual GPU

- Compatibility
  - Any physical GPU
  - Any guest driver stack
  - Adjustable capability exposure
  - No direct access to GPU memory

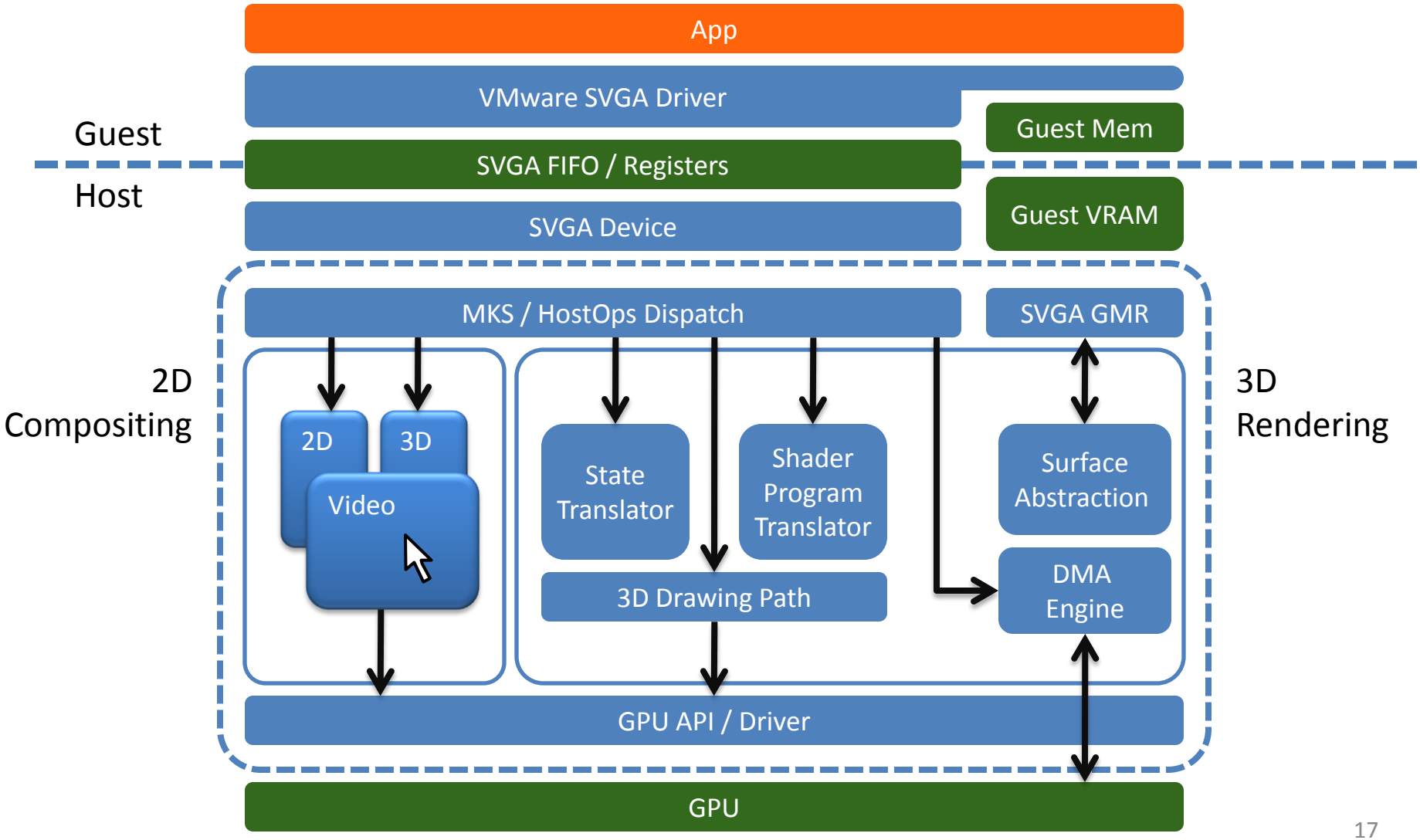


- Efficiency
  - Flexible guest memory management
  - Few copies
  - Asynchronous rendering

VMware SVGA II



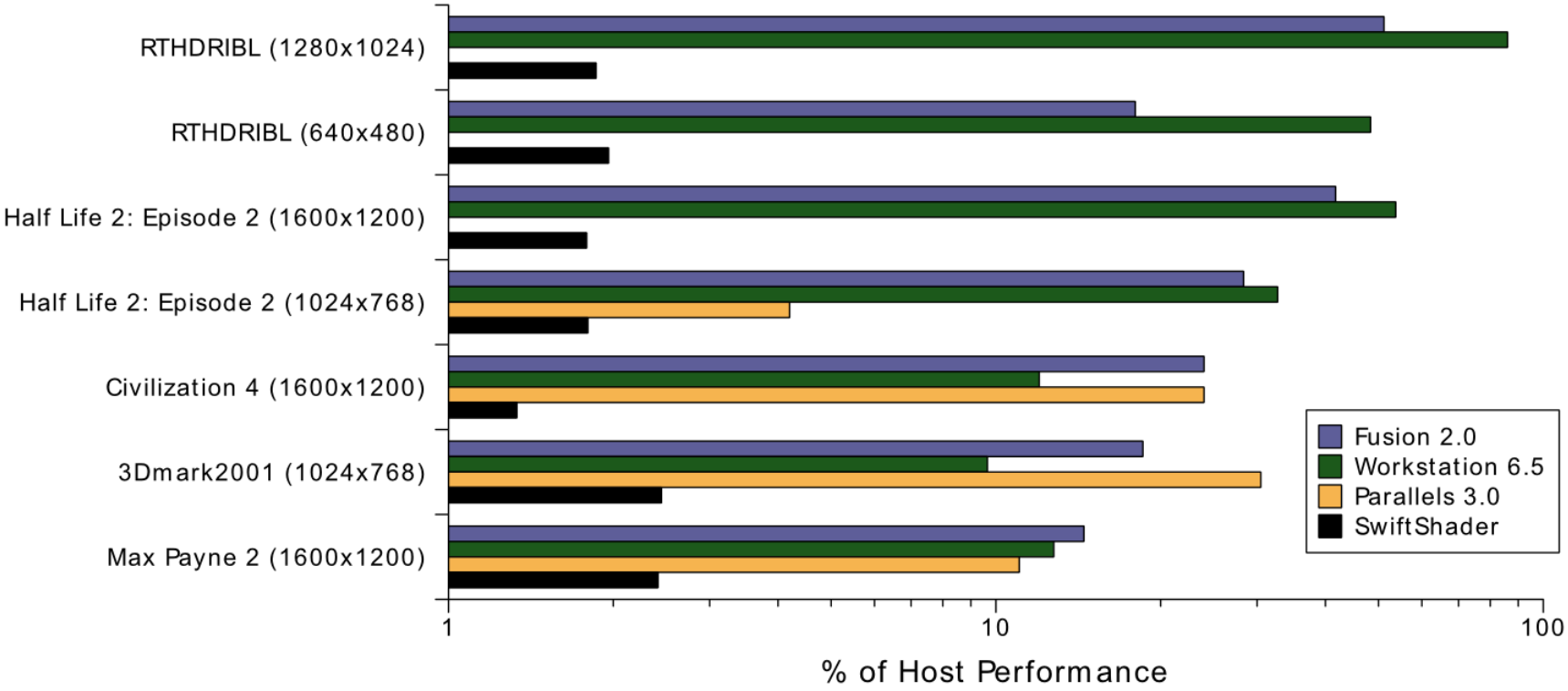
Virtual Graphics Stack



Evaluation

- Applications
- Microbenchmarks
- VMware Fusion 2.0, VMware Workstation 6.5, Parallels Desktop 3.0, SwiftShader
- Mac Pro, 8-core 2.8 GHz
- ATI Radeon HD2600

Application Benchmarks



Application	Resolution	FPS
RTHDRIBL	1280×1024	22
RTHDRIBL	640×480	27.5
Half Life 2: Episode 2	1600×1200	22.2
Half Life 2: Episode 2	1024×768	32.2
Civilization 4	1600×1200	18
Max Payne 2	1600×1200	42

Summary

- GPU Virtualization is an important problem
- Room for improvement in implementation completeness and performance...
- But we can already run interactive apps that could never be virtualized before
- Virtual GPU preserves portability + isolation

Future Work

- Pass-through techniques
  - Fixed and Mediated
  - Can be complementary to Virtual GPU
- Continued improvements
  - Performance and functionality
  - At all layers of driver stack
- Virtualization-aware GPU benchmarks

Questions?

- micah@vmware.com