Considerations for the deployment of GPUs in virtual desktop environments

Tennessee Higher Education Information Technology Symposium 2018

Jeremy Ey
Systems Administrator, ITS, Tennessee Technological University
vExpert 2018, VCIX6-DTM, VCP6-DCV

@kayakerscout

Questions

1. Have users currently running native or web applications on tablets, laptops, and desktops?
2. Currently have some form of desktop or application delivery?
   - Gave up on delivery of an application due to poor user experience?
3. Currently have GPU hardware in some desktop or application delivery hosts?
   - Currently have GPU hardware in all desktop and application delivery hosts?

Outline

- ~~Introduction~~
- Motivation
- Solutions
- Evaluation
- Conclusions

Motivation

Samsung Galaxy S9

Samsung Galaxy S9 image from https://news.samsung.com/us/gallery-samsung-galaxy-s9-s9-plus/

Qualcomm® Adreno™ 630 Visual Processing Subsystem

Snapdragon 845 Mobile Platform

- Open GL ES 3.2, Open CL 2.0, Vulkan, DirectX 12
- Ultra HD Premium video playback and encoding @ 4K (3840x2160) 60fps, 10bit HDR, Rec 2020 color gamut
- Slow motion HEVC video encoding of either HD (720p) video up to 480fps or FHD (1080p) up to 240fps
- H.264 (AVC), H.265 (HEVC), VP9, DisplayPort over USB Type-C support
- eXtended Reality (XR)
- Room-Scale 6DoF with simultaneous localization and mapping (SLAM)
- 2400x2400 @ 120 FPS per eye
- Adreno Foveation: multiple technology advancements for multi-view, tile-based foveation with eye-tracking and fine grain preemption

Intel NUC image from
https://newsroom.intel.com/news/intel-launches-powerful-intel-nuc-smallest-vr-capable-system-ever/

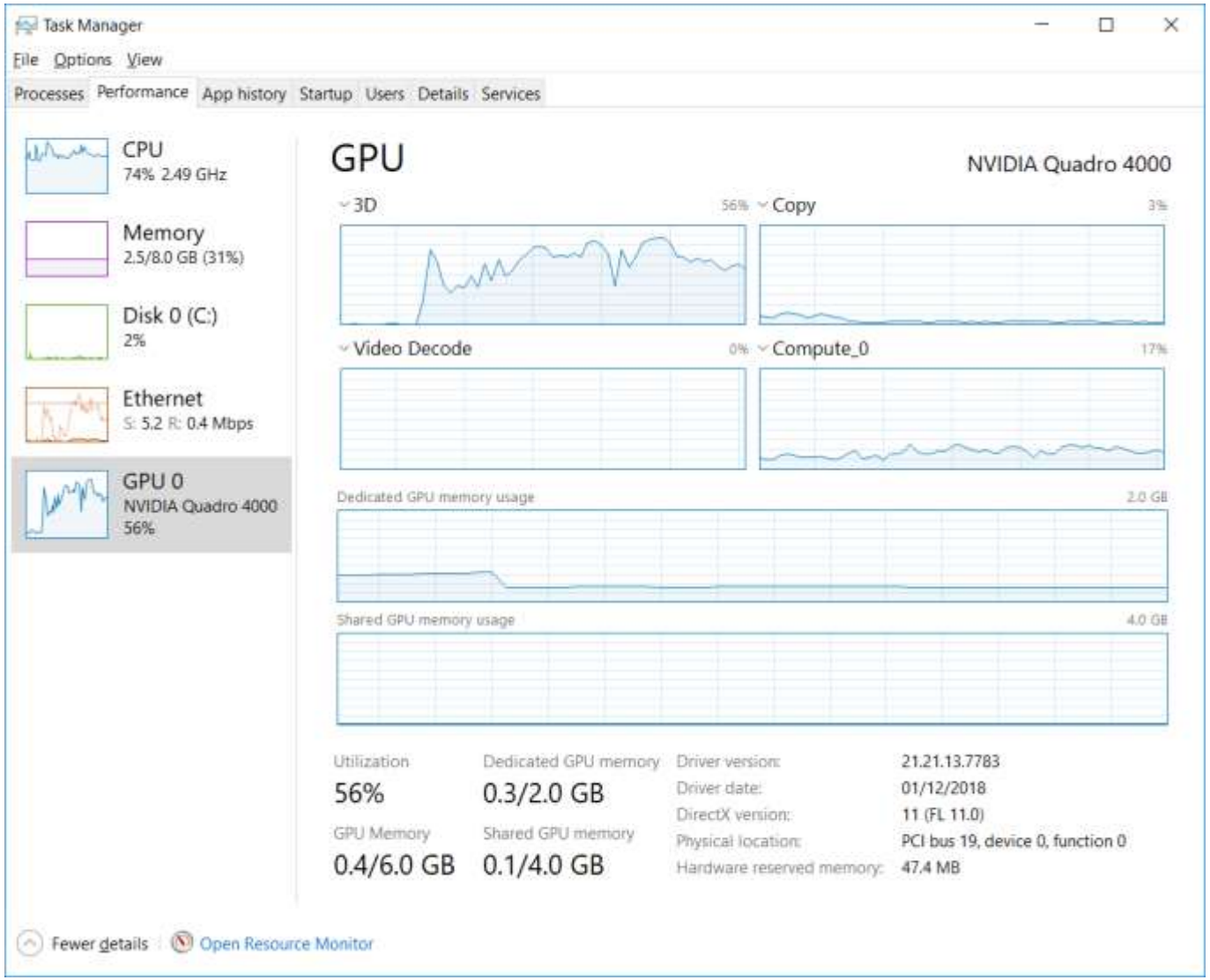
Intel NUC

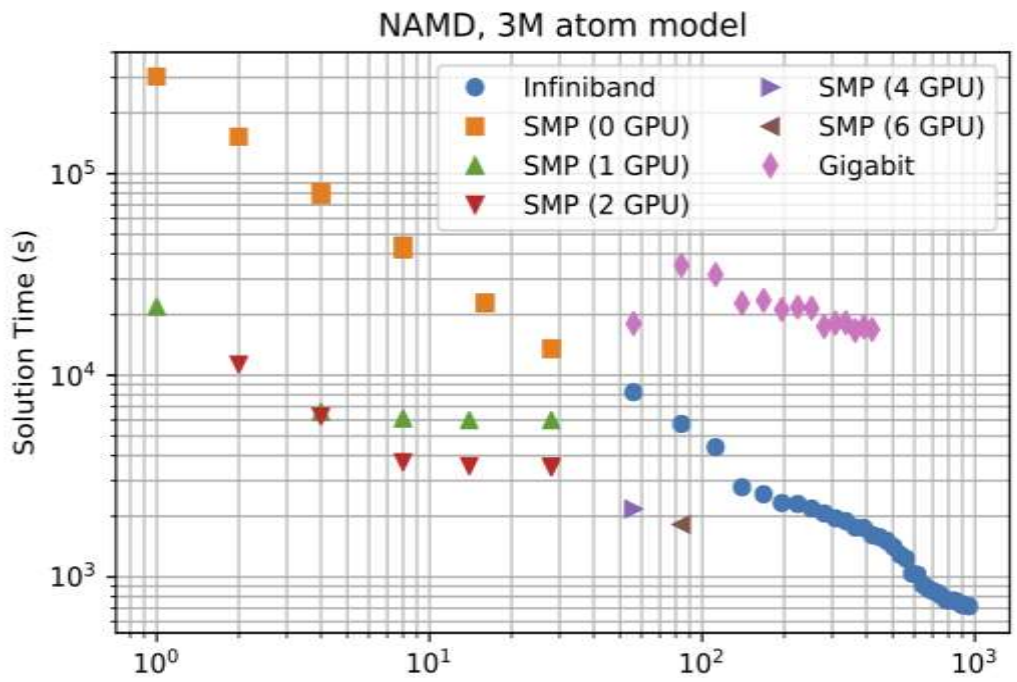
NUC8i7HVK
Radeon™ RX Vega M GH graphics

NUC8i7HNK
Radeon™ RX Vega M GL graphics

Windows 10

pictured:
Windows 10 1709 Task Manager
(GPU graphs required a GPU installed)





using 1-2 GPUs per GPU-enabled server results in runtimes equivalent to using 3-5 times as many non-GPU-enabled servers

video still from https://www.youtube.com/watch?v=wYl8Vv-qDfI

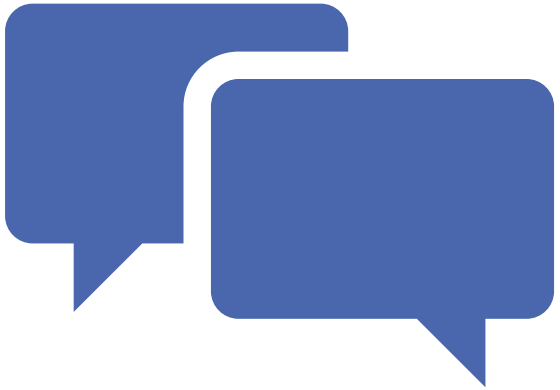
Motivation

Organizational Needs

User Expectations

Questions

1. Have formal/informal expectations for the performance of your application delivery environment?
2. Have expectations for the performance/capabilities of your application delivery environment based on devices outside that environment?
3. For the “optimizations” that are applied in your environment, which ones improve the user experience? Which ones sacrifice user experience to something else?



Solutions

Direct Graphics Adaptor (vDGA)

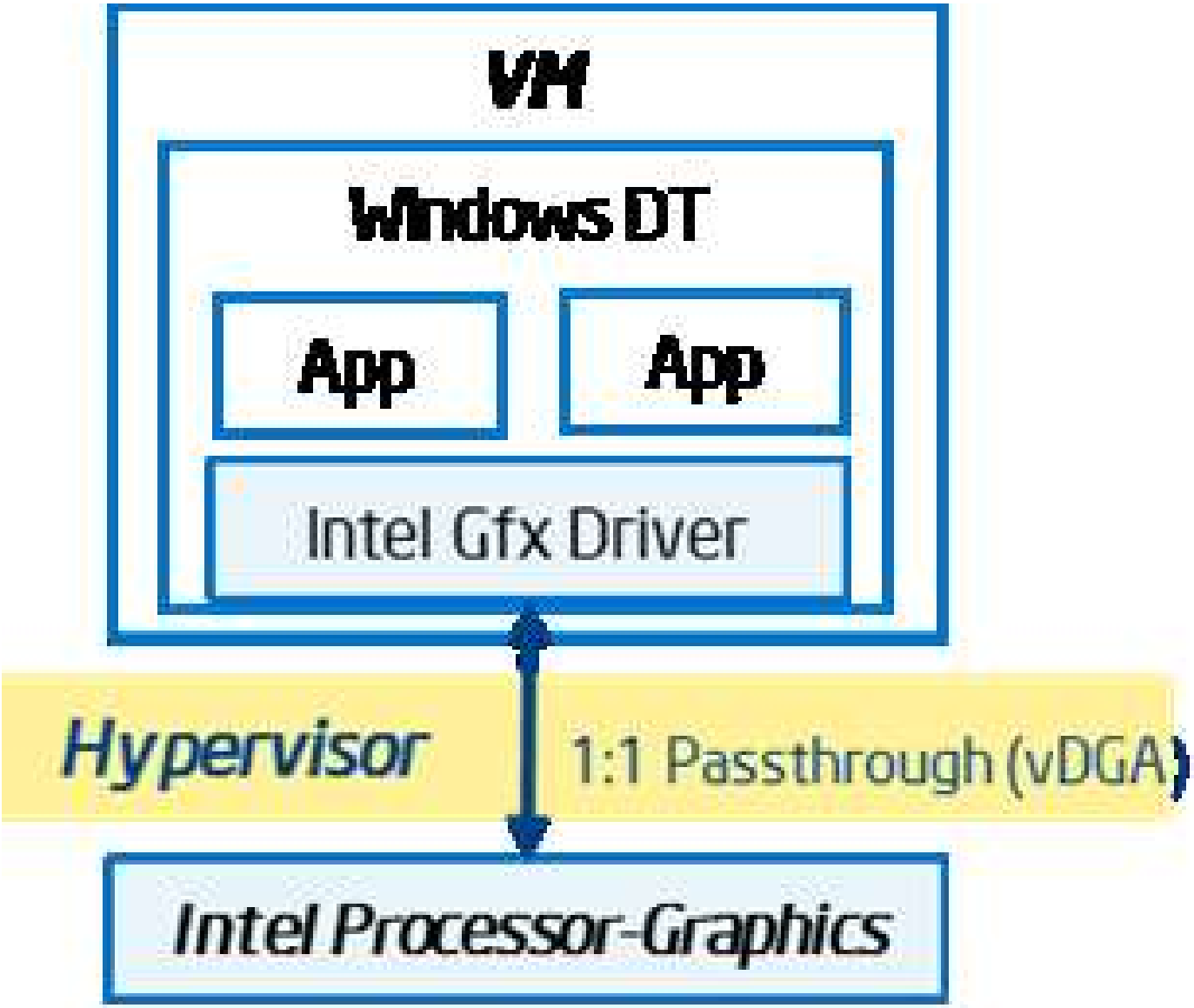


Image from Intel: https://01.org/blogs/2014/intel%C2%AE-graphics-virtualization-update

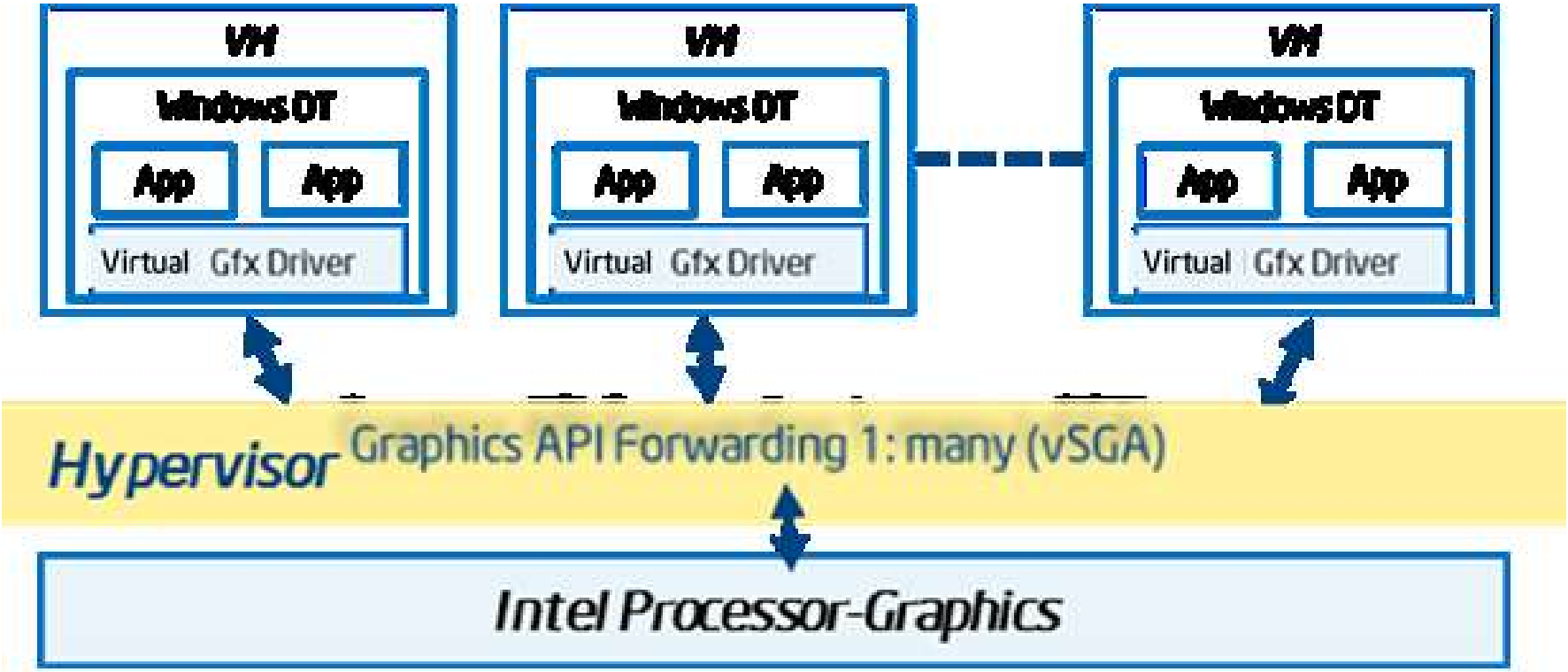


Image from Intel: https://01.org/blogs/2014/intel%C2%AE-graphics-virtualization-update

Virtual Shared Graphics Acceleration (vSGA)

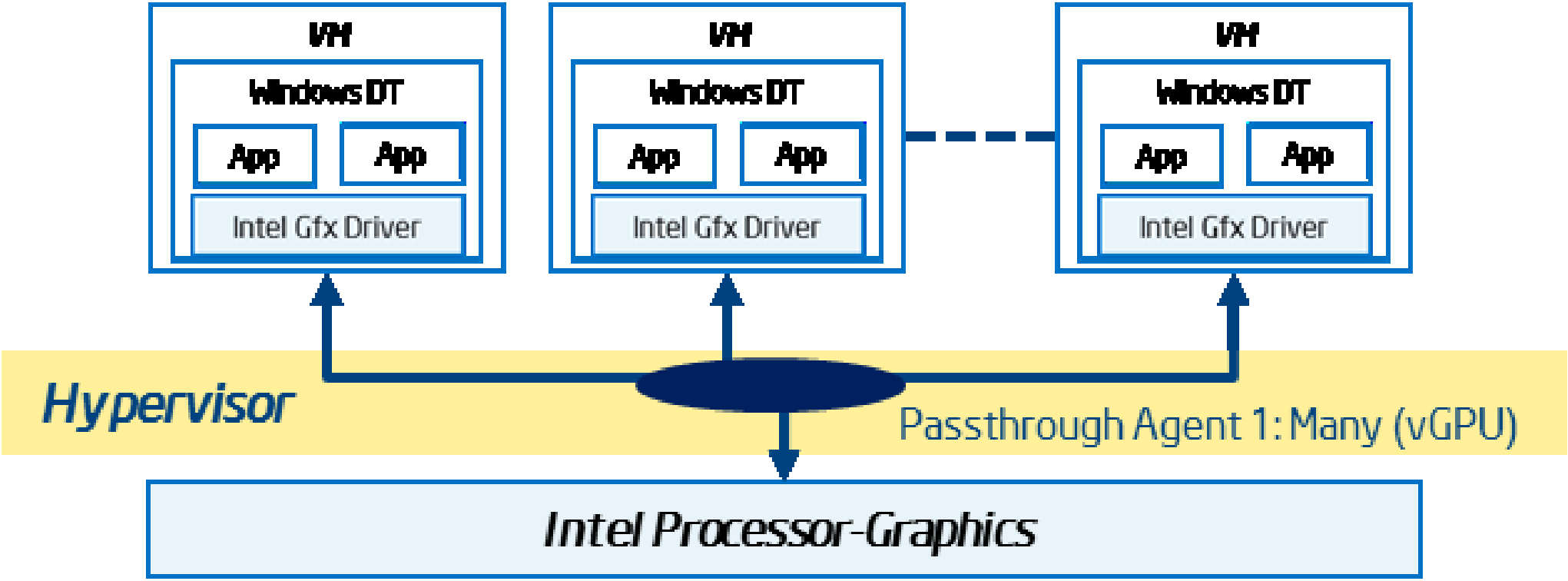


Image from Intel: https://01.org/blogs/2014/intel%C2%AE-graphics-virtualization-update

Virtual GPU

Solution Options

AMD

Intel

NVIDIA

Other Factors

Standard basics (storage, networking, processor/memory)

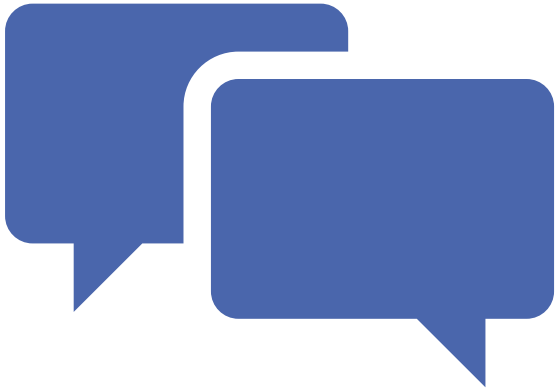
Virtual machine density

Remoting protocol (h264)

Provisioning system

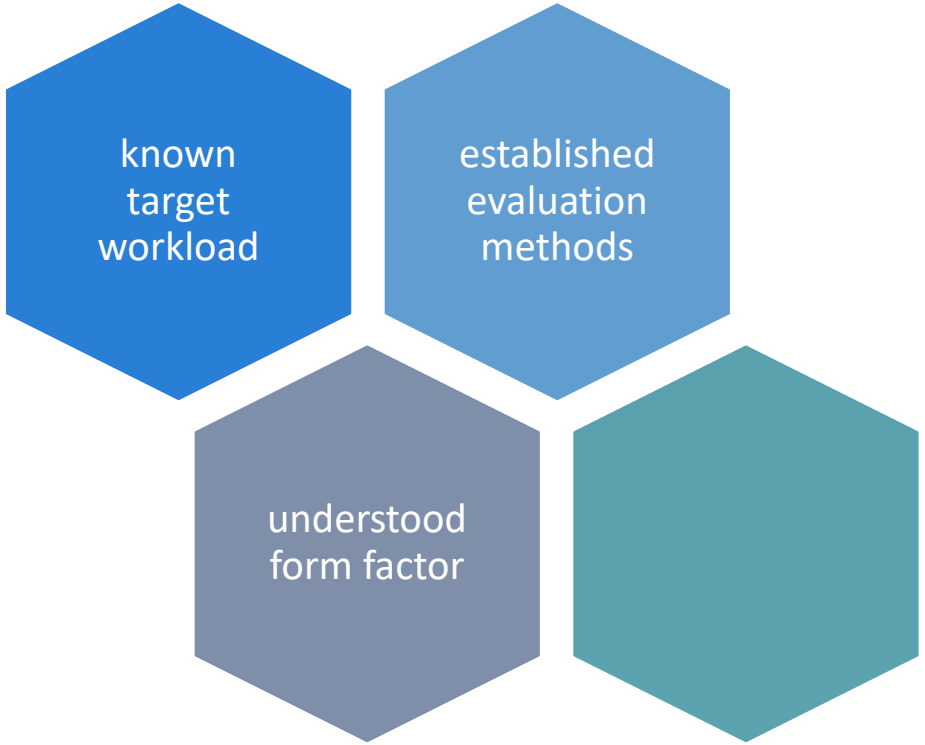
Questions

1. Can your existing/planed deployment practices accommodate deploying one or more of these solutions?
2. Are your density requirements/expectations compatible with one or more of these solutions?
3. Are your workloads static enough that you can fix your requirements for the full live of your hardware purchase?
4. How responsive can you be to changes in your workload requirements?



Evaluation

TOOLS AND TECHNIQUES



Physical Systems

Performance Monitor

File Action View Window Help

Performance
- Monitoring Tools
  - Performance Moni
- Data Collector Sets
- Reports

100 90 80 70 60 50 40 30 20 10 0

1:34:43 AM 1:34:56 AM 1:35:06 AM 1:35:16 AM 1:35:26 AM 1:35:35 AM 1:35:45 AM 1:35:55 AM 1:36:05 AM 1:36:1... 1:36:21 AM

Add Counters

Available counters

Select counters from computer:

<Local computer> Browse...

Generic IKEv1, AuthIP, and IKEv2
GPU Adapter Memory
Dedicated Usage
Shared Usage
Total Committed
GPU Engine
Running Time
Utilization Percentage
GPU Local Adapter Memory

Instances of selected object:

Search

Add >>

Added counters

Counter	Parent	Insta...	Computer

Remove <<

Show description

OK Cancel

Last 31.906 Average 20.148 Minimum 4.694 Maximum 78.549 Duration 1:40

Show	Color	Scale	Counter	Instance	Parent	Object	Computer
☑		1.0	% Processor Time	_Total	---	Processor Information	\\FRUITLOOPS

nvidia-smi

```
[root@esxi:~] nvidia-smi
Fri Aug 11 17:56:22 2018
+-----------------------------------------------------------+
| NVIDIA-SMI 390.42     Driver Version: 390.42              |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla M60           On   | 0000:85:00.0     Off |                  Off |
| N/A   23C    P8    23W / 150W |     13MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M60           On   | 0000:86:00.0     Off |                  Off |
| N/A   29C    P8    23W / 150W |     13MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P40           On   | 0000:87:00.0     Off |                  Off |
| N/A   21C    P8    18W / 250W |     53MiB / 24575MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Example nvidia-smi output from: https://docs.nvidia.com/grid/latest/grid-vgpu-user-guide/index.html

https://uberagent.com/d

List of Metrics • ub

File Edit View Favorites Tools Help

uberAgent

- Plugin load duration
- Plugin "load behavior" (state)
- Whether the plugin is installed per machine or per user

Related Splunk Events

Machine

Machine Performance

The following performance metrics are collected per machi

- CPU and RAM usage
- Relative CPU frequency
- GPU compute and memory usage
- GPU usage per engine (e.g. 3D, video decode; requires
- Kernel memory usage
- IOPS (read and write separately)
- Disk IO count, latency and volume (read and write sepa
- Disk utilization (percent disk time)
- Pagefile size and utilization
- Network utilization in percent
- Basic information about all active network interfaces
- Number of sessions, processes, threads, handles
- Idleness (how ready the machine is to go into power sa

SMB Client Performance

The following inventory metrics are collected per network s

- Share path
- IO count, latency, volume and IOPS (read and write)

Related Splunk Events

https://www.lakesides...

Introducing the NVIDIA Gra...

File Edit View Favorites Tools Help

Lakeside

Get Started

Introducing the NVIDIA Graphics Assessment

August 24, 2016 by Ben Murphy

When NVIDIA first announced their groundbreaking approach to introducing accelerated graphics to virtualization they began bridging one of the last gaps in making sure that users get the best possible experience with virtual desktops and applications. Building on their success NVIDIA has introduced newer, Maxwell™ based GRID cards NVIDIA that go even further to create a rich graphical experience for users and decrease complexity for IT administrators looking to optimize the visual experience of their users.

The evolution of the GRID solution set coupled with their new software means that more users than ever can take advantage of graphical acceleration. This couldn't come at a better time given the rise of advanced media usage in enterprise. As an increasing number of organizations begin to explore making their virtualized environments even better many want to explore what their current graphical needs are and plan for the future with NVIDIA.

This is why Lakeside Software has partnered with NVIDIA to develop a totally free graphical assessment hosted on Azure to deliver a detailed series of reports leveraging the SysTrack workplace analytics platform. Available for a 30-day period for up to 500 systems at no cost, this

100%

vRealize Ope...

ze Operations for
GRID

ger for VMware End-User Computing

rizon & Published Apps 6.5, including NVIDIA
generally available!

ealize Operations for Horizon

on customers have come to rely on VMware
d-to-end visibility into the health, performance
, virtual desktop and application
vices. And we're about to pump up those

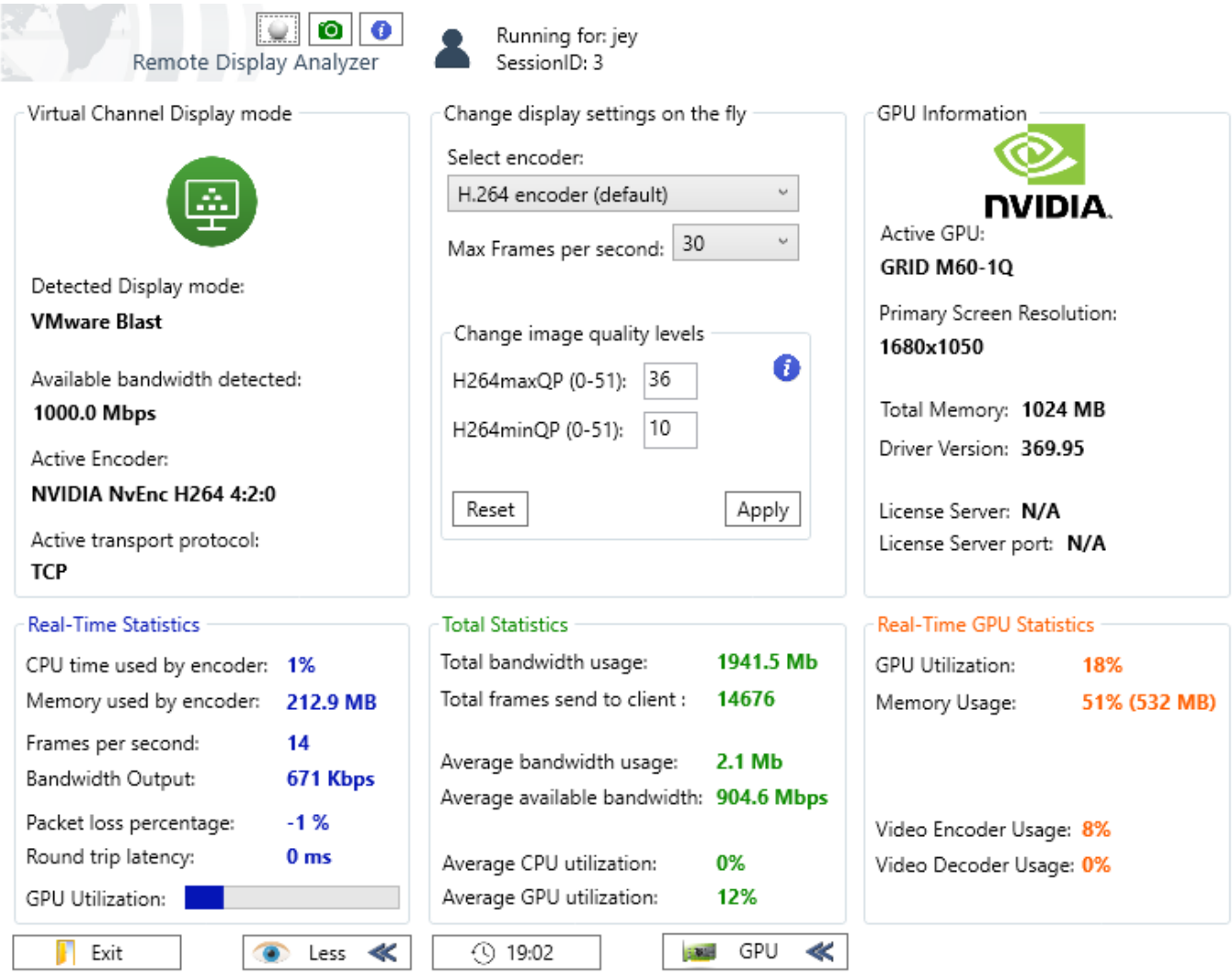
e and NVIDIA have been working closely to
he vRealize Operations for Horizon solution.
tors to gain greater insights and trend data on
of NVIDIA GPU products—including NVIDIA
ter Workstation Software—deployed with

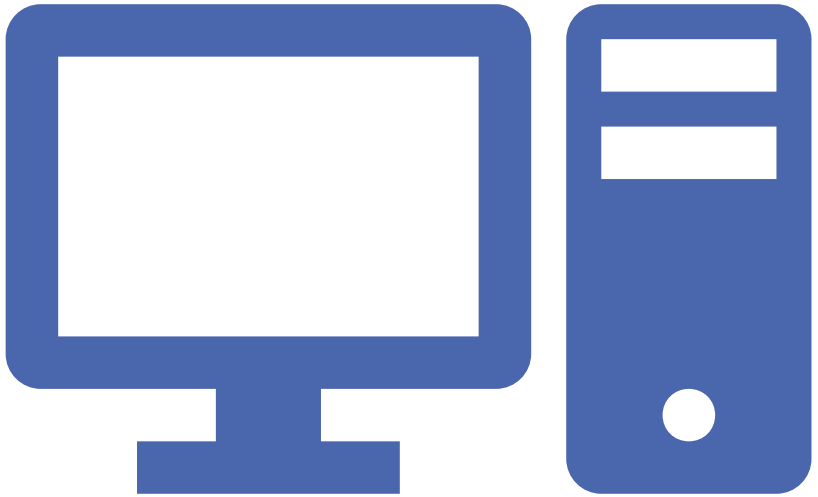
e Operations for NVIDIA management pack,
will expand to support the following use

100%

Remote Display Analyzer

https://www.rdanalyzer.com/





Workloads

Test using the applications your users use

Make sure you test using those applications the same way your users use them

Test under the same conditions your users work under (application and user concurrency)

LoginVSI screenshots: https://loginvsi.com/products/login-vsi

Testing at Scale

Workload Benchmarks

PCMark

Application Specific
examples: SolidWorks, Matlab

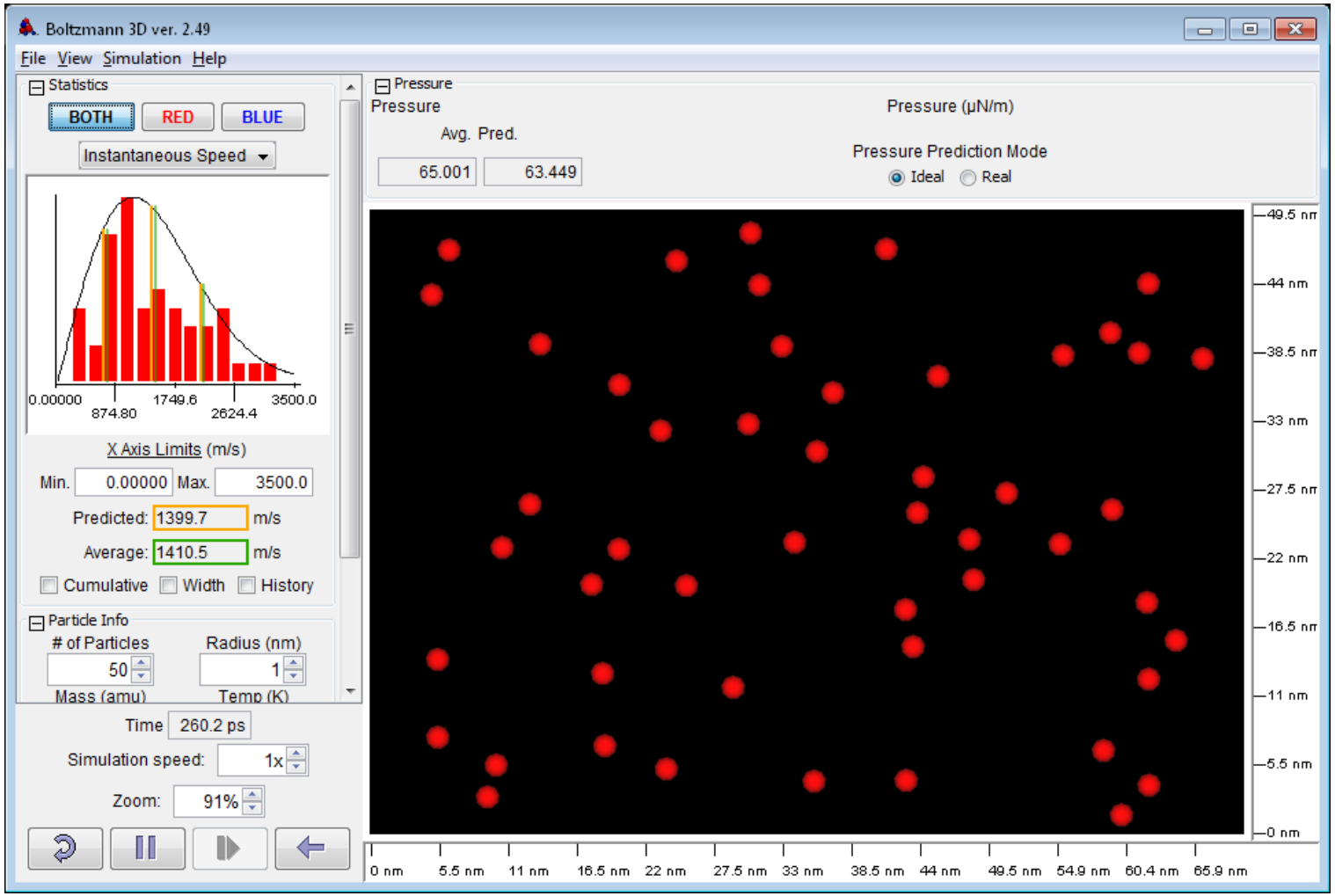
SPEC – GWPG
examples: SPECviewperf, SPECwpc

Unigine Heaven Benchmark

Not a Fix for Every Problem

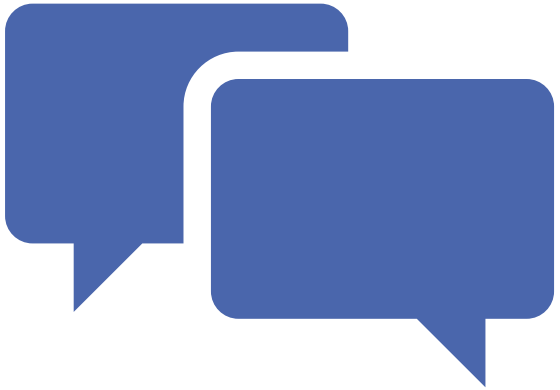
Boltzmann 3D – Java Application

Javaw.exe –Dsun.java2d.noddraw=true



Questions

1. How well do you know your users?
2. How well do you know your users' applications?
3. How well do you know how your users use their applications?
4. How can you capture and reproduce your users' workloads?



Conclusions

Community

State of EUC Survey
https://vdilikeapro.com/

GeekOut365 & TeamRGE Live Event
https://www.brianmadden.com/geekout365/playlist/5499158995001-5714546978001

Don't go it alone

- user groups
- online communities

Follow Up

Jeremy Ey
email:
twitter: @kayakerscout
blog: https://quirkyvirtualization.net

Discussion