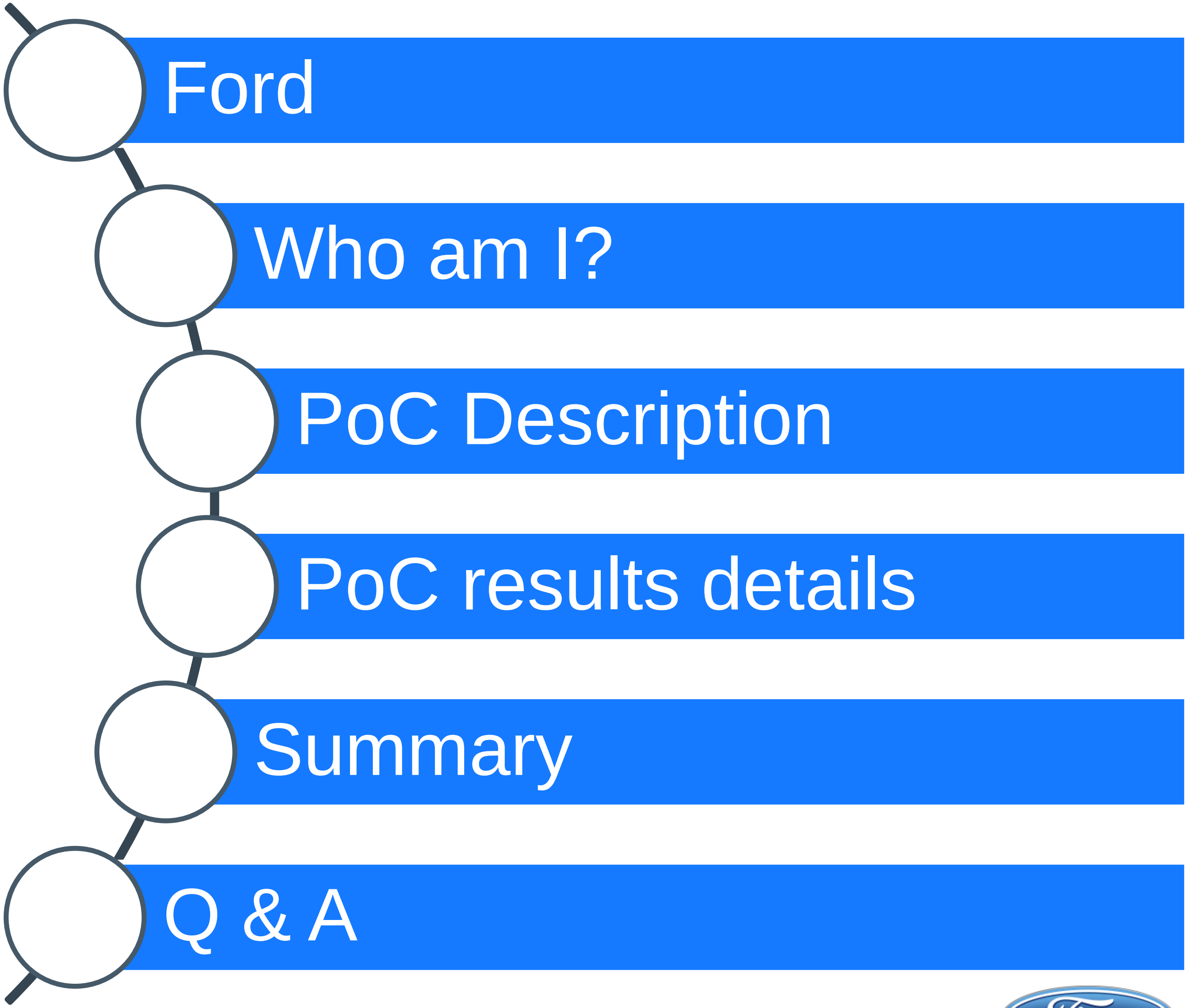
HIGH RESOLUTION GRAPHICS DESKTOP VIRTUALIZATION POC

Chip Charnley
Technical Expert – Client Technologies
Infrastructure Architecture

OCIO

Ford
Go Further



- Ford
- Who am I?
- PoC Description
- PoC results details
- Summary
- Q & A

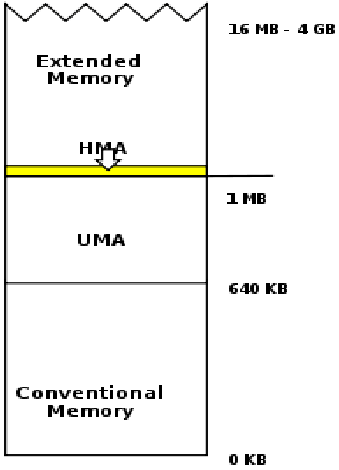
Go Further

- Ford Motor Company, a global automotive industry leader based in Dearborn, Mich.,
- Manufactures or distributes automobiles across six continents.
- With about 186,000 employees and 65 plants worldwide, the company's automotive brands include Ford and Lincoln.
- The company provides financial services through Ford Motor Credit Company.

For more information regarding Ford and its products worldwide, please visit www.corporate.ford.com.

Who am I?

- Chip Charnley
- Worked at Ford since 1989
- Currently part of the Infrastructure Architecture group as the Client Technologies Technical Expert.
- First client technology was to build a Windows 3.1 IP based client using QEMM and FTP software in 1995.
- Spent a significant portion of the last 5 years focused on implementing Virtual Desktops at Ford.



Ford and HRG VDI

- Ford has been investigating High Resolution Graphics (HRG) Desktop Virtualization off and on for close to 8 years although the reasons have morphed over time.
- Investigation started as a way to improve software delivery/patch management especially when 1000's of machines globally had to be updated in one weekend.
- Most recently the driving reasons have been protection of intellectual property when dealing with joint venture and supplier access to Ford data sources and global performance to centralized data centers (latency driven performance and data load timing).

Ford and HRG VDI

- Two prior Proofs of Concept (POC) ended in failure due to unacceptable infrastructure cost per user and unacceptable end user experience due to high latency to global user locations.
- The Siemens Team Center Visualization demo in the Synergy 2013 keynote convinced me that the technology was finally available to solve most of Ford's end-user experience/ performance issues at a price point that the business would find acceptable.
- So I sought out those who had been involved in prior efforts or had expressed recent interest to participate in a new PoC in 2014

2014 PoC

- Ford partnered with Citrix and Cisco
  - PoC architecture was driven by already in place non-HRG VDI architecture @ Ford.
- PoC consisted of two phases
  - Benchmark XenApp (XA) and XenDesktop (XD) on various infrastructure configurations
  - Execute end-user testing on a target configuration to determine if the end-user experience was acceptable
- The rest of this presentation will primarily focus on the benchmark phase of the PoC

PoC team

- The core technical team was made up of myself and Engineers from Citrix and Cisco.
- Cisco also provided
  - project management support
  - architecture design support
- Ford also provided
  - network design/services to connect the PoC environment to the Ford production network
  - project management support
  - SME’s to test end-user experience w/ selected apps

PoC Hardware BoM

- Design based on existing VDI implementation @ Ford
  - Replaced blades with 4x C-240M3s Servers (2x E5-2680v2 2.8GHz CPUs) to support nVidia GRID cards
    - One used for virtualized management servers
    - 3 used for various XA and XD configurations
  - Added nVidia GRID GPU cards
    - 4 NVIDIA GRID K2
    - 2 NVIDIA GRID K1
  - Cisco rack network w/ Nexus switches and Cisco firewalls to connect to the Ford Intranet

PoC Infrastructure Software BoM

- Design based on existing VDI implementation @ Ford but most software versions were updated to adequately support the nVidia GRID cards
  - Utilized XenServer 6.2 SP1 (Hotfixes XS62ESP1003, XS62ESP1005,XS62ESP1008) vs. ESX
  - XenDesktop 7.5
  - XenApp 7.5
  - Windows Server 2012R2
  - Cisco UCS 2.2.2c

Why Benchmark?

- No good industry data on optimal XA/XD hardware configuration w/ nVidia GRID required to meet Ford requirements.
- Determine for XA
  - Which is better, physical or virtual servers
  - what combination of RAM, CPU, & GPU (K1/K2) is likely to provide the best user density and performance mix
- Determine for XD
  - what combination of RAM, CPU, GPU (K1/K2) and vGPU profile is likely to provide the best user density and performance mix

Benchmark Methodology

- Utilized RedWay's RedTurbine demo program to generate load.
  - –RedTurbine was chosen because:
    - Reasonable facsimile of a CAD workload
    - Ability to set to run continuously
  - –Heavy load = continuous run
  - –Medium load = Random start, random run-time, random restart intended to more accurately simulate 'real' usage
  - –Utilized 4 laptops to open end-user sessions
  - –1 dedicated to a single user experience
  - –3 used to generate as many sessions as required to meet a given benchmark target.

Benchmark Methodology (cont.)

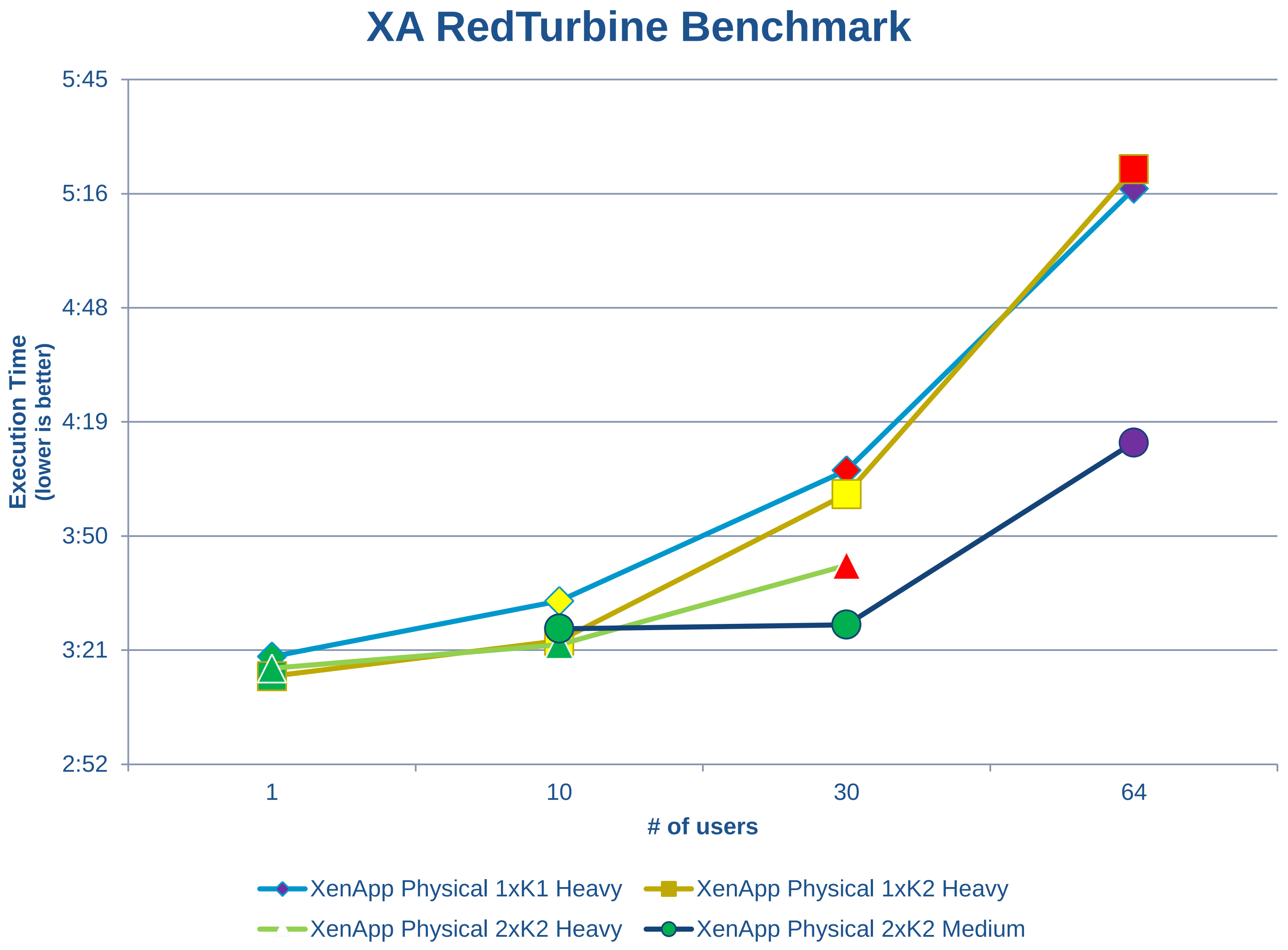
- Key performance data points collected
  - A qualitative (subjective human) assessment of the visual performance (Excellent, Good, Degraded, Poor).
  - Time required for a full cycle of the RedTurbine demo. (manually timed with a stopwatch)
  - Physical CPU utilization %
  - Physical GPU utilization %
  - vCPU utilization %
- Above recorded for 1, 8/10, 16, 30 & 64 users
- Human SMEs from 3 target applications provided feedback on end-user experience.

XenApp Benchmark results

Results were surprising:

- A physical server implementation with uncontrolled allocation of multiple GPUs to the processes significantly outperformed multiple VM instances with single dedicated GPU per instance.
- Benchmarking shows 30 users per server as acceptable and 64 users as unacceptable
- Optimum user density will always depend on the application configuration and usage patterns
- The ability of this configuration to send commands to the GPU appears to be the bottleneck as GPU utilization never exceeded 77% regardless of user density.

XenApp Data Charts



Except as noted below and the one user scenario, the GPUs were generally utilized evenly across all GPUs. In all scenarios except the one user scenario, the GPU utilization never went above 77%!

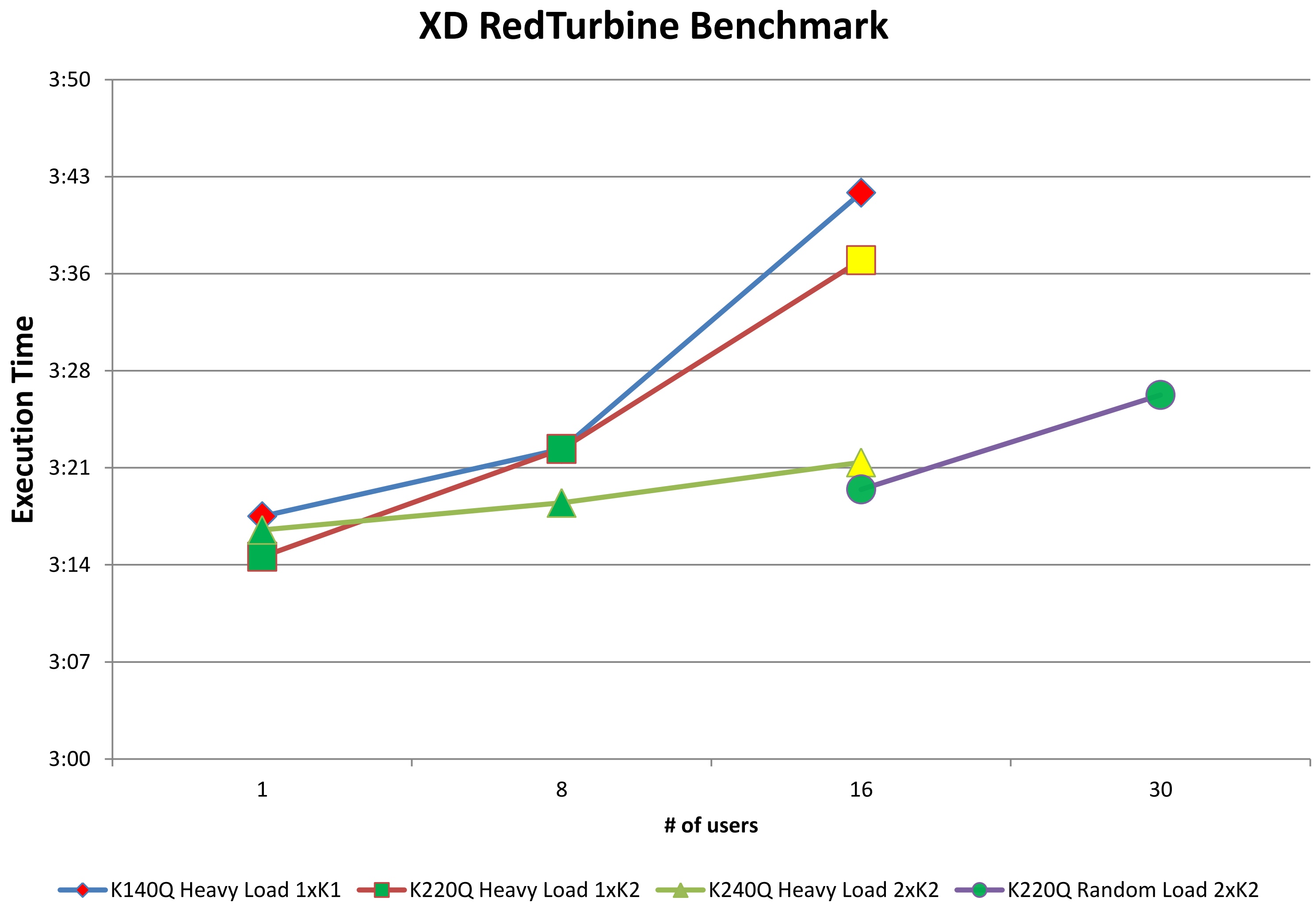
The Physical 2xK2 Heavy @ 30 users went south due to something new - For the first and only time we saw CPU queuing and lopsided utilization of the GPUs. There were also indications that NUMA transitions were part of the problem.

Markers are color coded to Qualitative assessment of video display
Green = Excellent
Yellow = Good
Red = Degraded
Purple = Poor

XenDesktop Benchmark results

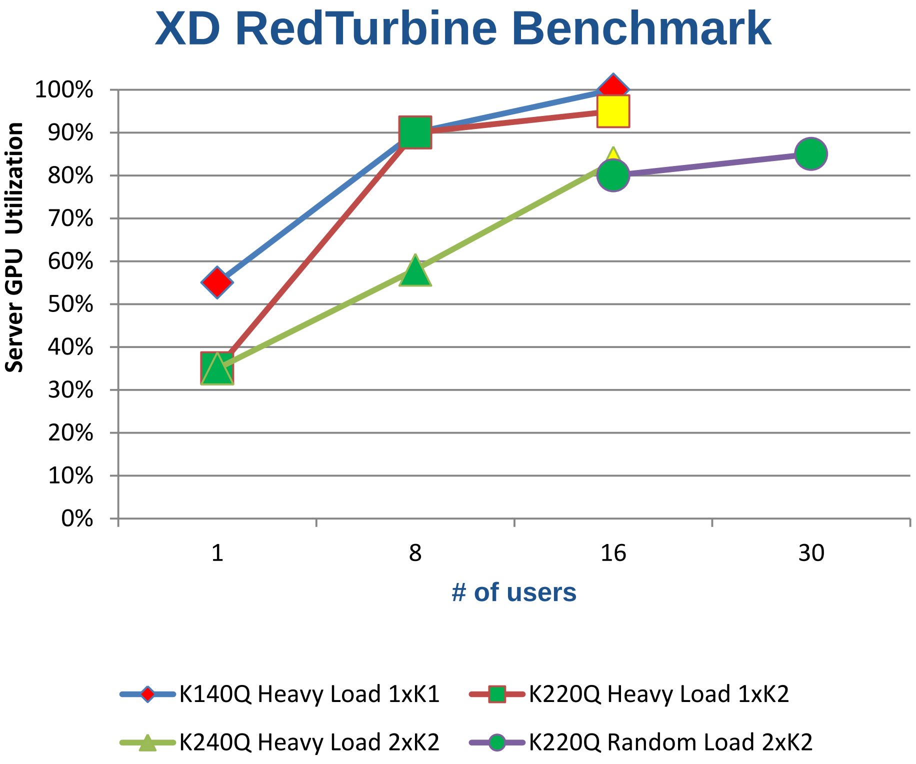
- Initial tests quickly showed that K1 cards were not going to provide the level of user experience desired at the CAD/CAE level
- Given the server CPU and RAM available, the K220Q vGPU profile with dual nVidia GRID K2 cards provided the best utilization of all resources (CPU, RAM and GPU)
  - Acceptable performance across 30 users at moderate usage
  - Max user density of 32 users for K220Q could not be tested due to RAM limitations
    - Testing of other vGPU profiles at max user density suggests that increasing RAM to get to 32 users for K220Q might result in lower end-user experience (due to slower memory bus speed) and significant additional cost with a lot of RAM left unused.
- SME end-user experience performance perception showed differences between the RedTurbine results and the testing of the 3 target applications

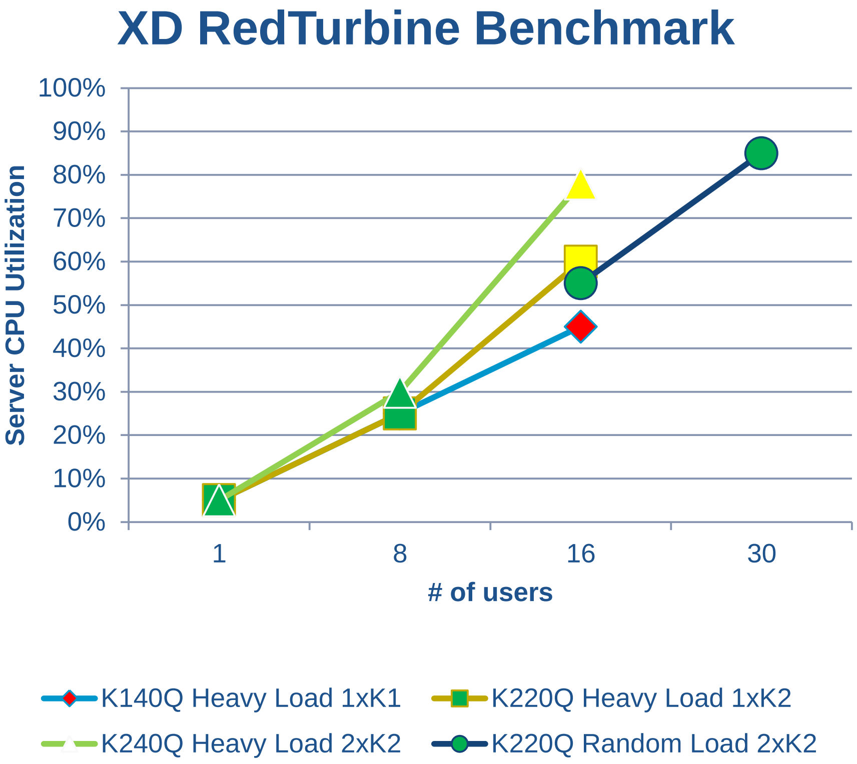
XenDesktop Data Charts



Markers are color coded to Qualitative assessment of video display
Green = Excellent
Yellow = Good
Red = Degraded
Purple = Poor

XenDesktop Data Charts





Markers are color coded to Qualitative assessment of video display
Green = Excellent
Yellow = Good
Red = Degraded
Purple = Poor

Shows that the hardware is optimized with a K220Q vGPU profile and 30 users.

XD Latency Testing

		225ms	250ms	275ms	300ms	325 ms	350ms	375ms
CAD App1		🟩	🟩	🟩	🟩	🟩	🟨	🟥
CAE App1		🟩	🟩	🟩	🟩	🟨	🟨	🟥
CAD App2		🟩	🟩	🟩	🟩	🟨	🟨	🟥

- Green is acceptable performance
- Yellow is potentially acceptable performance
- Red is unacceptable performance

- This information is not from the RedTurbine benchmark but from actual SME (human) visual assessment of end-user experience.
- This is interesting because the industry information that could be found in early 2014 suggested to the PoC team that the top end acceptable latency would be in the 200ms-250ms range.

Summary

- Both XD and XA appear capable of providing acceptable performance while increasing end user productivity by as much as an order of magnitude (reduced data load times).
- Both appear to have user density (per user cost) that will be justifiable for the targeted Ford business cases.
- While the benchmark points to a K220Q profile as being the most cost effective, other testing clearly indicates higher capability/cost user profiles will be required to support some use cases.
- The impact of each application and the end user usage pattern on user density cannot be over stated.

Q&A

Chip Charnley
ccharnle@ford.com