REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03/27/2023	Conference Briefing	

4. TITLE AND SUBTITLE

Hybrid Computing Capability Evaluation

5a. CONTRACT NUMBER

W56KGU-18-D-0004

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. Sacha Panic; Brian Parrish; Bruce Gorski; Aidan Buffum; Kaeden Le; David Laratta;

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730

8. PERFORMING ORGANIZATION REPORT NUMBER

PRS-23-0786

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

United States Army (ACC-APG)

10. SPONSOR/MONITOR'S ACRONYM(S)

ACC-APG

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

Preprint to NVIDIA GTC

14. ABSTRACT

Overview of the Hybrid Computing Capability Evaluation work in progress. Phase I, the current scope of work, the goal is to investigate how different QoS configurations of Nvidia GeForce NOW affect the QoS for cloud gaming at scale. At present the team is in the process of finalizing the building blocks needed to conduct the envisioned evaluations. We are creating a generic approach that can be used to facilitate evaluation of the performance of any training simulation software used on cloud gaming infrastructure – at any scale.

15. SUBJECT TERMS

Training Technologies; Computing Methodologies (General); hybrid computing; synthetic training environments; cloud gaming

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Susan Carpenito
				9	19b. TELEPHONE NUMBER *(Include area code)* 781-271-7646

Hybrid Computing Capability Evaluation

2023/03/09

Dr. Sacha Panic (apanic@mitre.org)
Brian Parrish (bparrish@mitre.org)
Bruce Gorski
Aidan Buffum
Kaeden Le
David Laratta

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Hello! I am Sacha Panic, and in the next few minutes I will be informing you of the Hybrid Computing Capability Evaluation work in progress, done by a team of people at The MITRE Corporation, in collaboration with the Institute for Creative Technologies at the University of Southern California, as well as Cole Engineering Services.

Background: MITRE Prior Work

Synthetic Training Environment (STE) Cross Functional Team (CFT) 5G “Cloud Gaming” Experiment

- Key Observations
  - Streaming Training Simulation Software (TSS; e.g. Bohemia Interactive VBS3) video at 1080p, 60 frames per second, 18 Mbps, provided a smooth gaming experience
  - Gameplay experience acceptable until >0.1% packet loss and/or 180.42 ms end-to-end latency

- Selected Recommendations
  1. Investigate with multi-player (echelon size) condition
  2. Investigate tradeoffs between software efficiency, hardware specifications, power consumption, cooling, and network latency in context of the gameplay experience and training outcomes
  3. Repeat the Experiment using the latest Training Simulation Software (TSS) to provide better data and better observations/benchmarks
  4. Experiment with automated video stream bandwidth modulation to fit available network resources
  5. Evaluate Augmented/Virtual Reality (AR/VR) streaming

MITRE

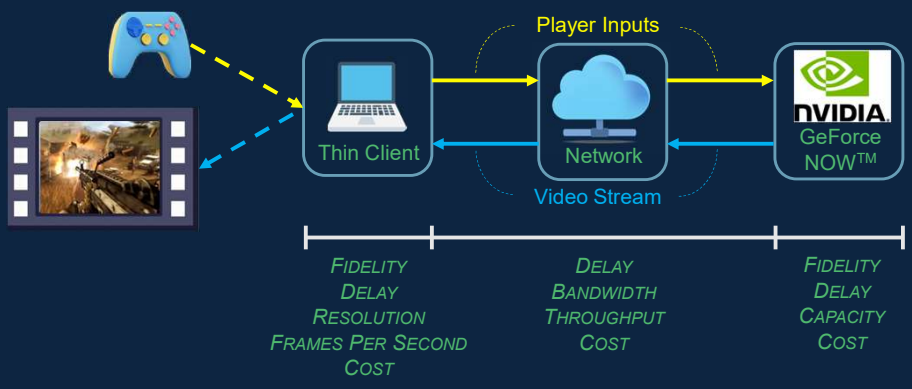
Cloud gaming involves rendering 3D game content on a server, and streaming the video to a device operated by the end user. One of the key enabling technologies is 5G communication

Recent work by the MITRE corporation showed that streaming video at 1080p resolution, 60 fps (which amounts to 18mbps) provided a smooth gaming experience, and that the gameplay experience was acceptable until at least 10% packet loss and/or 180msec end-to-end latency was introduced

Some of the selected recommendations were to conduct testing under operationally relevant conditions: more concurrent players, using current-gen training simulation software, and also consider automated video stream bandwidth modulation (similar to how Netflix changes the video resolution depending on network connectivity)

The on-prem use of Nvidia’s GeForce NOW capabilities enable evaluating these, other considerations, at scale.

Hybrid Computing Capability



Goal: Evaluate the (cost-)effectiveness of NVidia GeForce NOW™ in combination with Thin Clients to deliver training to large number of concurrent trainees.

MITRE

3

This diagram provides a functional depiction of cloud gaming, along with some of the associated uncertainties that are under consideration.

The user or trainee interacts with a thin client device, which has limited on-board computing or rendering capabilities. Their inputs are sent to the Nvidia GeForce NOW server, which updates the state of the training simulation software accordingly, before rendering a visual representation of it. The sequence of rendered frames is then streamed to the thin client in video format.

Each of the steps involved in this process contributes to the delay that the user experiences between pressing a key on the keyboard or gamepad, and seeing the associated effect reflected in the video stream.

While end-to-end latency can be minimized, and visual fidelity can be maximized, by improving the enabling infrastructure – this comes at a cost.

This raises the question: when large numbers of concurrent users are involved, would it be more cost-effective to operationalize using cloud gaming using Nvidia GeForce NOW, or using thick clients (i.e. gaming laptops with Nvidia RTX GPU's?

Towards Guiding Research Questions

- Typical “cloud gaming” metrics[1]:
  - Quality of Service (QoS): the ability to dependably run high-priority applications and associated network traffic under limited network capacity
  - Quality of Experience (QoE): ties together user perception, experience, and expectations to application and network performance[2]
- Competing stakeholder perspectives:
  - Service provider: the (cost-)efficiency of QoS configuration schemes is crucial
  - Trainee: the QoE, as affected by QoS, is crucial
  - Trainer: configuring QoS so that resulting QoE does not affect training outcome[3], is crucial
- What is an acceptable compromise?

1. Metzger, F., Geißler, S., Grigorjew, A., Loh, F., Moldovan, C., Seufert, M. and Hoßfeld, T., 2022. An Introduction to Online Video Game QoS and QoE Influencing Factors. IEEE Communications Surveys & Tutorials, 24(3), pp.1894-1925.
2. Fiedler, M., Hossfeld, T. and Tran-Gia, P., 2010. A generic quantitative relationship between quality of experience and quality of service. Ieee Network, 24(2), pp.36-4
3. Liu, S. and Claypool, M., 2022, April. The Impact of Latency on Navigation in a First-Person Perspective Game. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-11).

MITRE

Typically cloud gaming metrics fall in one of two categories. Quality of Service describe technical specifications or characteristics of the system, while Quality of Experience describe how end users perceive and affectively respond following interactions with it.

It is key to realize that in this particular situation there are competing stakeholder perspectives. What it comes down to is that the service provider (who owns all the infrastructure) wants it as cheap as possible, the user or trainee wants a positive and pleasant experience, and the trainer wants to be sure that the trainees are adequately prepared for the real-life situations that they may encounter in the (near) future.

It will be crucial then, to experimentally investigate what an acceptable compromise would be, in terms of configuration of cloud gaming infrastructure.

Guiding Research Questions

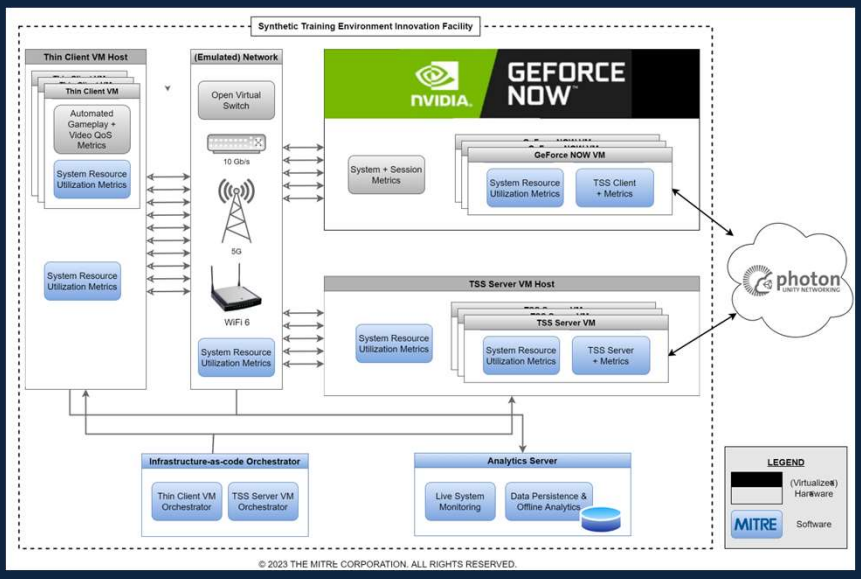
- Phase I (current scope)
  - How does Nvidia GeForce Now™ configuration affect the ability, in terms of QoS, to deliver "cloud gaming" based training to thin client devices?
  - In terms of operating cost, what is the break even point of using cloud gaming in combination with thin clients to deliver training to up to Battalion sized groups (100-1000 soldiers), vs. equipping trainees with thick clients/gaming laptops?
- Phase II (future work)
  - How do different Nvidia GeForce Now™ configurations affect QoE and training outcomes?

MITRE

In Phase I, the current scope of work, the goal is to investigate how different QoS configurations of Nvidia GeForce NOW affect the QoS for cloud gaming at scale. A technical report will ultimately inform a decision on the operational infrastructure that the sponsor should acquire in order to support computer-based training at scale.

While Phase I is purely focusing on QoS and, as you will see in the next slides, automated datacollection, Phase II will extend the investigation to include human-in-the-loop evaluations of QoE and training outcomes. As mentioned before, there is a tradeoff between "high fidelity smooth streaming experiences being expensive in terms of infrastructure" and "low fidelity experiences negatively affecting user experience, in-game decisions and behaviors, and as such, training outcomes".

Evaluation Architecture And Infrastructure



MITRE

6

This diagram shows the envisioned architecture of the scalable and configurable system we are developing to evaluate cloud gaming. Starting from the top right, the main system components are:

- The Nvidia GeForce NOW servers. These host virtual machines that run customized operationally relevant Training Simulation Software, which has been instrumented with our metrics collection tools. We will also capture system and session resource utilization data from virtual and physical server hardware.

- To the left of Nvidia GeForce NOW is the Emulated network. We can optionally emulate different networking capabilities, investigating their impact on the ability to facilitate cloud gaming at scale

- In the top left is the Thin Client infrastructure which will host automated game playing bots and collection of metrics related to video streaming. We will again capture system resource utilization data from virtual and physical server hardware.

- In the lower left corner is the Infrastructure-as-code orchestrator. This automates the process of instantiating and configuring virtual machines

- In the lower right corner is the analytics server. This is responsible for live monitoring and visualization of system status, as well as centralized collection and persistence of produced metrics

Key Architecture Components

Kafka[1]

An open-source distributed (high fidelity) data streaming platform

Prometheus[2]

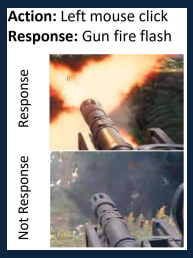
An open-source approach to (low fidelity) data provisioning

Grafana[3]

An open-source data monitoring & visualization solution

RIDE[4]

Rapid Integration & Development Environment. A Research & Development testbed accelerating DoD simulation technologies



DECAF[5]

DissEcting Cloud gAming perFormance using automated game playing bots, and AI/ML based measurement of end-to-end delay and video stream delivery performance

1. https://kafka.apache.org/
2. https://prometheus.io/
3. https://grafana.com/
4. https://ride.ict.usc.edu/
5. Iqbal, H., Khalid, A. and Shahzad, M., 2022, June. Dissecting Cloud Gaming Performance with DECAF. In Abstract Proceedings of the 2022 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (pp. 13-14).

MITRE

7

Our preferred approach is to (creatively) reuse existing building blocks as much as possible, as summarized in this slide.

In terms of metrics, we distinguish between low fidelity metrics used for live monitoring of system status, and high fidelity metrics for offline analysis. We use Kafka for distribution of high-fidelity metrics, which are centrally collected and stored to file, as well as in a database. Prometheus and Grafana are used for live monitoring of the system state using lower fidelity metrics. This helps us verify that during any given data collection run, the system is working as expected, with each component generating the anticipated data.

We collaborate with people at the Institute of Creative Technologies over at University of Southern California, who have developed RIDE – a framework built on top of the Unity 3D game engine, and which can expedite the creation of simulation-based training. This enables us to use current-gen technologies, molded into operationally relevant training scenarios – which will be instrumented with our metrics capture tools.

As mentioned before, the goal of the project is to investigate cloud gaming at scale. To mitigate the need to recruit hundreds of human participants, we use DECAF – an approach that uses automated game playing bots in combination with AI/ML based computer vision that uses keystroke or gamepad input logs, in combination with video stream snippets, to calculate end-to-end latency.

Current And Anticipated Impact

- Current impact
  - Technologies to transition: Distributed metrics collection for data collection and status monitoring, and the Analytics Server capabilities integrated in other work that The MITRE Corporation does for our sponsors; with initial interest from industry & academic collaborators
- Anticipated impact
  - A flexible and scalable approach to facilitate evaluation of arbitrary TSS performance on Nvidia GeForce NOW™ cloud gaming infrastructure, and how QoS, QoE, and training outcomes are affected
  - Three on-site data collection events in FY23 will result in a technical report containing actionable recommendations on cost-effectiveness of acquisition & operationalization
  - Scientific report and/or presentation of findings at relevant conference or through publication in peer reviewed journal

MITRE

8

At present the team is in the process of finalizing the building blocks needed to conduct the envisioned evaluations. We are creating a generic approach that can be used to facilitate evaluation of the performance of any training simulation software used on cloud gaming infrastructure – at any scale,

The results from three on-site datacollection events will be used to deliver a technical report, as well as present some of our findings through appropriate scientific channels.

Some of the software components that we have developed under this effort, have proven to be valuable for other stakeholders internal or external to The MITRE Corporation. These initiatives have benefited from our work by integrating analytics server and distributed (low and high fidelity) metrics collection into their tools.

This concludes the overview of our work in progress.

On behalf of the team at The MITRE Corporation, I'd like to thank our collaborators at the Institute of Creative Technologies at the University of Southern California, and Cole Engineering Services Incorporated. And thank you for your interest in our work in progress.